UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Sistemas de recomendación en el contexto gastronómico: elaboración y enriquecimiento de un dataset de recetas de cocina

Máster Universitario en Ingeniería Informática

Autor: MÉNDEZ LÓPEZ, Francisco

Tutor: DÍEZ RUBIO, Fernando
Departamento de Ingeniería Informática

Septiembre, 2018

Resumen

El uso de sistemas de recomendación en Internet está teniendo una popularidad creciente y son utilizados en diferentes tipos de aplicaciones de la Web. En el contexto de la gastronomía, Internet ha contribuido a la diversidad en la cultura gastronómica de las personas, permitiendo que existan diversos tipos de aplicaciones orientadas a la gastronomía, como por ejemplo las redes sociales de cocina. Los sistemas de recomendación también han aterrizado en el ámbito de la gastronomía, posibilitando la existencia de aplicaciones en la Web que realizan sugerencias de platos de comida a los usuarios.

En este trabajo se realiza un proceso completo de minería de datos, que abarca desde la adquisición de datos hasta la puesta en funcionamiento de modelos analíticos predictivos. El trabajo comienza con una revisión del estado del arte en los sistemas de recomendación en el contexto de la gastronomía, principalmente en la búsqueda de datos de recetas de comida que sean totalmente públicos y que cumplan una serie de requerimientos determinados. Ante la complejidad para encontrar este tipo de datos, se ha decidido realizar un proceso de extracción de información, a partir de una fuente de datos de recetas de comida, escogida como resultado de un análisis previo de las diferentes fuentes de datos que hay disponibles. La extracción de datos constituye la parte central de este trabajo, para lo cual se ha desarrollado un software automático que extrae los datos de manera incremental y automática. Este programa ha permitido la obtención de los datos requeridos, coleccionando un total de 362 usuarios, 12151 recetas y 495210 revisiones con rating asociado. El trabajo restante abarca el desarrollo y evaluación de dos sistemas de recomendación, basados en contenido, y utilizando los datos que se han obtenido previamente mediante el software de extracción de datos.

Palabras clave

Gastronomía, receta de comida, sistemas de recomendación, extracción de datos, Minería de Datos, Recuperación de Información, *scraping*, *crawling*, Big Data, procesamiento distribuido.

Abstract

The use of recommender systems on the Internet is experiencing a growing popularity as lots of applications across the Web are offering personalized content. In the context of gastronomy, the Internet has contributed positively to the diversity of the gastronomic culture around people, allowing the existence of many different gastronomy-oriented web applications on the Internet, such as cooking-oriented social networks. Recommender systems have also left a mark on gastronomy, since we can find on the Internet lots of applications suggesting food recipes to Internet users.

In this work a complete data mining process is performed, starting from data acquisition and ending with the testing of various predictive models. This work starts with the revision of the state-of-the-art trends in gastronomy-oriented recommender systems, mainly focused on the search of any available public dataset about cooking recipes, given a set of restrictions. Given the difficulty at finding this kind of data, we have decided to develop a program to extract that data from a previously analyzed data source that met a set of requirements. The data extraction process is the main part of this work, and in order to get the required data we have developed a software that obtains that data in an incremental and automatic manner, using various scraping and crawling techniques. Through this program we have been able to collect a set of 362 users, 12151 cooking recipes and 495210 revisions with associated rating. The remaining part of this work is oriented to the development and evaluation of two content-based recommender systems that use the data previously gathered by the data extraction software developed.

Keywords

Gastronomy, cooking recipe, recommender systems, data extraction, Data Mining, Information Retrieval, *scraping*, *crawling*, Big Data, distributed processing.

Índice de contenidos

Índice de tablas

Índice de ilustraciones

Capítulo 1

Introducción

En este primer capítulo se realiza una introducción al trabajo, indicando la motivación que ha permitido formalizarlo, los objetivos que se pretenden alcanzar y una visión general de cómo está organizado el documento.

1. Motivación

Desde hace unos años la web ha evolucionado hacia escenarios en los que los usuarios cada vez generan mayor cantidad de contenidos, tal y como ocurre en la Web 2.0. Como consecuencia de esta evolución los distintos servicios disponibles en Internet ofrecen contenidos personalizados orientados a mejorar su calidad y la experiencia del usuario, basándose en sus gustos y preferencias, que quedan reflejados en la interacción con los contenidos de la Web.

En esta línea los contextos de la cocina y la gastronomía también se han visto involucrados en los nuevos avances de documentación y personalización de contenidos. Actualmente los sistemas son capaces de recoger múltiples datos que reflejan las preferencias de las personas en el mundo culinario. Son numerosas las aplicaciones en las que los usuarios pueden publicar contenido relacionado con la cocina. De igual modo, el número de estudios e investigaciones científicas realizados en torno a la minería de datos sobre gastronomía crece rápidamente. Las investigaciones y los avances que éstos posibilitan se orientan hacia la recomendación de recetas, planes de comida u opiniones acerca de restaurantes, tratando de mejorar lo máximo posible la experiencia de usuario, teniendo en cuenta la información que se dispone del mismo. La minería de datos y la inteligencia computacional se aplican a un mundo donde la función vital de la nutrición se convierte en pasión por la cocina, donde la calidad de la comida cobra importancia y, además, los gustos de las personas son muy variados. Por estas razones son por las que el trabajo se ha orientado hacia este ámbito de mejora de la experiencia de usuario a través de la personalización.

En este contexto, para la formalización de este trabajo ha sido realizada una revisión sobre el estado del arte en sistemas de recomendación orientados a la cocina, así como la disponibilidad de conjuntos de datos públicos con los que realizar experimentos e investigaciones. Los resultados de esta revisión evidencian, en primer lugar, que la mayoría de las investigaciones realizadas en este campo han utilizado datos con un conjunto reducido de características, muchas veces limitados a valores numéricos o datos que no incluyen un registro de la actividad y gustos de los usuarios de manera extensa. En segundo lugar, y a pesar de las numerosas investigaciones realizadas en este ámbito, se ha observado la carencia de conjuntos de datos con la suficiente cantidad de información acerca del usuario y de

recetas de cocina que sean totalmente públicos y accesibles; en algunas ocasiones los autores han utilizado datos que han sido obtenidos ad-hoc para sus estudios, pero en pocos casos son públicos y no cumplen los requisitos explicados en el primer caso de características limitadas.

Como consecuencia de lo anterior y ante la carencia de datos públicos que satisfagan requisitos de completitud y extensión suficientes, se ha decidido realizar este trabajo, con el objetivo de elaborar un dataset que palíe, en la medida de lo posible, alguna de las deficiencias observadas. Además, con el dataset elaborado se va a evaluar el comportamiento de distintos modelos de recomendación, partiendo del conocimiento adquirido en el estudio previo ya realizado.

2. Objetivos

De acuerdo con la motivación expuesta anteriormente, el objetivo general del trabajo es la elaboración de un dataset que contenga información tanto de recetas de cocina como de sus autores y demás usuarios que hayan interactuado con ellas de algún modo y, posteriormente, realizar procesos de elaboración e interpretación de los datos, con el fin de extraer y generalizar conocimiento en el mundo de la cocina.

Para la consecución de dicho objetivo general se proponen los siguientes objetivos parciales:

Objetivo O1

Revisión del estado del arte sobre recomendación en el ámbito de la cocina y la gastronomía. Aplicaciones, fuentes de datos, algoritmos, resultados, etc.

Objetivo O2

Obtención y extracción de datos de recetas de cocina y de usuarios de los sitios web más utilizados. Generar una versión preliminar de un conjunto de datos consistente y amplio, que contenga información diversa acerca de las recetas, de los usuarios y de las relaciones entre éstos.

Objetivo O3

Pre-procesamiento de los datos con el fin de construir una versión final de un dataset con información estructurada para cada usuario y receta, en el que se haya eliminado información irrelevante, así como también se haya añadido información y características relevantes, deducidas a partir de otras propiedades. La versión final de este dataset debe intentar ser lo más amplia y consistente posible.

Además, como parte de este objetivo se pretende enriquecer los datos obtenidos con nuevas características inferidas a partir de datos de otras fuentes. En particular se plantea añadir al dataset la *dificultad* de la receta así como su *salubridad* (entendida como el grado en que una receta es saludable por sus ingredientes o por su forma de cocinado). Se pretende analizar la hipótesis de si, añadiendo estas características, se pueden mejorar los resultados de recomendación a partir del dataset base.

Objetivo O4

Procesamiento de los datos. Utilización del dataset generado de usuarios y recetas para realizar recomendaciones, incorporando algunos de los algoritmos ya existentes relacionados con este ámbito.

Objetivo O5

Evaluación y contraste de los resultados obtenidos, pudiendo reflejar el conocimiento obtenido a partir de los datos mediante herramientas de visualización.

3. Estructura general del documento

El documento se organiza de la siguiente manera:

En el capítulo 1 se realiza una introducción al trabajo realizado, detallando la motivación como punto de partida y los distintos objetivos que se han planteado.

Una vez se han formalizado los objetivos del trabajo, se ha realizado una revisión del estado del arte, tanto en la disponibilidad de conjuntos de datos públicos como en sistemas de recomendación en el contexto gastronómico, cuyo resultado puede encontrarse en el capítulo 2.

Tras la revisión del estado del arte se ha comenzado con la búsqueda de datos gastronómicos. Este proceso se ha descrito en el capítulo 3, y consiste en la definición de las características y restricciones del conjunto de datos que se desea obtener, y con ello la selección de la fuente de datos más apropiada desde la cual obtener los datos.

La primera parte técnica del trabajo, detallada en el capítulo 4, ha consistido en la generación de un programa de extracción de datos, de la fuente de datos previamente seleccionada, y que contempla todas las fases de desarrollo del software, incluyendo análisis (de la fuente de datos escogida) y captura de requerimientos del dataset a generar, diseño del programa y del modelo de datos, implementación del software y detalles y resultados de la ejecución.

En el capitulo 5 se definen los pasos realizados para llevar a cabo la aplicación de sistemas de recomendación al dataset que se ha generado, incluyendo el diseño de los diferentes algoritmos que se han utilizado y el detalle de los resultados obtenidos.

Por último, en el capítulo 6, se realiza una reflexión general sobre el trabajo realizado, incluyendo las dificultades encontradas y los hitos conseguidos.

Introducción

Capítulo 2

Estado del Arte

El proceso de revisión ha comenzado con la búsqueda de datos libres que contengan información de usuarios y recetas de cocina. Esta búsqueda de datos libres se ha realizado teniendo en cuenta un conjunto de restricciones a nivel general que permitirán importar los datos a los modelos básicos de sistemas de recomendación, que contemplan principalmente la existencia de recetas y usuarios y que contengan información básica de cada receta y usuario, así como valoraciones y revisiones o comentarios que los usuarios realizan a las recetas.

1. Datos de gastronomía

Una vez definidos estos requisitos generales del dataset se ha comenzado con el proceso de investigación, en el cual han sido revisados un total de 40 artículos, detallados en la Tabla 1, que o bien versan sobre temas relacionados con sistemas de recomendación en el mundo culinario o bien mencionan o utilizan datos de recetas de cocina. Además, también han sido incluidos aquellos artículos que han sido referenciados por otros que satisfacen alguna de las dos condiciones anteriores, dando lugar a una variopinta lista de temas que van desde la utilización de grafos de ingredientes para detectar comunidades y con ello sus respectivos sabores [1], hasta reconocimiento de imágenes de recetas de cocina [20]. Para cada artículo se ha extraído una serie de características reflejadas en la Tabla 1.

Estado del Arte

Artículo	Tema del artículo				Dataset							
	Análisis de datos	Recomendación de recetas	Diseño de software	Transformación de datos	Dataset público	Obtención ad hoc	Contiene imágenes/vídeos	Datos dietéticos	Contiene recetas	Contiene datos nutricionales	Contiene ratings/revisiones	Contiene logs
(Ten et al., 2012)		X				X			X	X	X	
(Ahn et al., 2011)	X					X			X			
(Kinouchi et al., 2008)	X						X		-	-	-	
(Shidochi et al., 2009)	X			X		X	X		-	-	-	
(Hashimoto et al., 2008)			X			X			X			
(Malmaud et al., 2014)			X	X	X				X			
(De la Torre et al., 2009)				X	X		X	X	-	-	-	
(Jermsurawong et al., 2015)				X	X				X			
(Ueda et al., 2011)		X				X			X		X	
(Weber et al., 2016)	X					X		X				
(Achananuparp et al., 2016)	X	X				X		X				
(Wagner et al., 2014)	X					X			X			
(West et al., 2013)	X							X				X
(Said et al., 2014)	X					X			X		X	
(Harvey et al., 2012)	X		X			X			X	X		
(Harvey et al., 2013)		X				X			X	X	X	
(Kusmierczyk et al., 2015)	X					X			X	X	X	
(Freyne et al., 2010)		X				X			X		X	
(Chung, 2012)	X					X			X			
(Wang et al., 2015)	X		X		X		X		-	-	-	

Estado del Arte

Artículo	Tema del artículo				Dataset							
	Análisis de datos	**Recomendación de recetas**	**Diseño de software**	**Transformación de datos**	**Dataset público**	**Obtención ad hoc**	**Contiene imágenes/vídeos**	**Datos dietéticos**	**Contiene recetas**	**Contiene datos nutricionales**	**Contiene ratings/revisiones**	**Contiene logs**
(Freyne et al., 2011)		X				X			X		X	
(Forbes et al., 2011)		X				X			X	X	X	
(Wagner et al., 2014)	X					X			X			X
(El-Dasuky et al., 2012)		X	X		X				X	X		
(Berkovsky et al., 2010)		X				X			X			
(Safreno et al., 2013)	X					X			X			
(Kuo et al., 2012)		X				X			X			
(Aberg, 2006)		X			?	?						
(Greene, 2015)					X				X			
(Nedovic, 2013)	X								X			
(Shinsuke et al., 2012)				X		X			X			
(Trevisiol et al., 2014)	X	X				X					X	
(Yamakata et al., 2013)	X			X		X			X			
(Wesley et al., 2016)	X					X		X				
(Tasse et al., 2008)			X	X	X				X			
(Zhang et al., 2008)	X			X	?	?			X			
(Wang et al., 2008)		X		X	X	X			X			
(Ueda et al., 2014)		X							X		X	
(Muller et al., 2012)	X					X			X			
(Shunsuke et al., 2015)	X					X			X			
Número de artículos (total: 40)	21	14	6	9	**8**	**27**	4	5	**30**	6	10	2

Tabla 1: Lista de artículos revisados y su clasificación según los temas que abarcan y las características de los datos que utilizan. 'X' implica que el artículo cumple la condición, '-' si la condición no aplica y '?' si no se ha podido determinar a partir del artículo.

En vista de lo datos reflejados en la tabla se tiene lo siguiente:

- Un porcentaje muy bajo (8 sobre 40) de artículos han utilizado datos públicos.
- 27 artículos de 40 han realizado procesos para obtener datos según los requerimientos de su investigación.
- 30 de 40 artículos utilizan datos de recetas de cocina, de los cuales 24 de ellos han obtenido sus datos ad hoc para su estudio.

Según los puntos anteriores, podemos observar la dificultad para obtener datos públicos y un alto porcentaje de artículos en gastronomía que han obtenido datos ad hoc para su estudio.

Las distintas fuentes de datos que los autores han utilizado son las siguientes:

- Bases de datos para reconocimiento de imágenes. Entre éstas están la de la *Universidad Carnegie Mellon* [7], que se trata de una base de datos construida a partir de la grabación de personas cocinando recetas y capturando así datos de acelerómetros, audio, imágenes y vídeo, y la base de datos de la *Universidad Pierre and Marie Curie* [20], que alberga imágenes de 100.000 recetas clasificadas en 101 categorías.
- Logs de buscadores. En algún artículo, como por ejemplo en [13], se han analizado registros de buscadores web con el fin de inferir las preferencias del usuario a partir de su comportamiento en la navegación web.
- Sitios web de cocina y de dieta. La mayor parte de las investigaciones han realizado extracciones de datos enfocadas al objetivo del estudio que querían abordar, obteniendo datos de sitios web populares sobre salud[1] [10] y sobre cocina en general[2]. En algunos casos estos datos sí cumplían con los requerimientos establecidos, pero en ningún momento los autores mencionan la disponibilidad de dichos datos.
- Recetarios de cocina. En otros estudios (como por ejemplo en [18]) los datos utilizados han sido obtenidos de recetarios online de cocina, como son los recetarios dietéticos de CSIRO[3]o *The British New Penguin Cookery Book*[4]. Del mismo modo que en el caso anterior, los datos que han utilizado no están disponibles y los recetarios mencionados no son totalmente públicos.

En una segunda búsqueda, orientada específicamente hacia sistemas de recomendación en gastronomía, se ha encontrado un artículo en el cual se realiza un extenso análisis del contexto que rodea a este tipo de sistemas y, entre otros temas, las fuentes de datos disponibles [43]. En él se hace mención a aquello que se ha visto reflejado en la anterior revisión realizada en este trabajo, y es la carencia de datos, para lo cual los diferentes investigadores recurren a una extracción ad hoc para su estudio de sitios web populares cuyos datos no pueden publicar por temas de legalidad. Los autores del artículo también hacen referencia a datos obtenidos del ya mencionado recetario *CSIRO'S Wellbeing Diet Book*, así como los sitios web Cookpad[5] y Yummly[6] con soporte académico para obtención de datos. Aunque estas dos últimas posibilidades resultan viables para este trabajo, implican realizar un proceso de extracción de datos que, aun siendo también factible, conlleva realizar un análisis para evaluar todas las posibles fuentes de datos, y así poder decidir cuál de ellas se adapta mejor a los requerimientos de este trabajo. Por último, se indica la existencia de un dataset disponible por parte del Instituto de Tecnología de Massachusetts (MIT) que provee de 1 millón de recetas y que, sin embargo, al carecer de perfiles de usuario y de interacciones

[1] https://www.myfitnesspal.com/

[2] http://allrecipes.com/
https://www.kochbar.de/
https://www.ichkoche.at/

[3] https://www.totalwellbeingdiet.com/

[4] https://www.penguin.co.uk/books/24596/the-new-penguin-cookery-book/

[5] https://cookpad.com/es

[6] https://www.yummly.com/

entre éstos y las recetas de comida, no es apto para ser utilizado en un contexto offline, que es precisamente uno de los escenarios que se quiere tratar como parte de este trabajo.

Tras este primer proceso de revisión se ha decidido incorporar al trabajo un proceso de obtención de datos como tema central y así poder construir un conjunto de datos estándar y completo que pueda estar disponible para otros trabajos.

2. Sistemas de recomendación en el contexto gastronómico

Paralelamente a la búsqueda de datos abiertos de cocina se ha realizado la revisión de otro de los contextos más importantes de este trabajo: sistemas de recomendación en el contexto gastronómico. En la actualidad existen numerosos sitios y aplicaciones web de cocina que tienen implementadas funcionalidades que sugieren a los usuarios nuevo contenido, basándose esencialmente en sus preferencias, reflejadas a través del contenido con el que han interactuado. Sin embargo este ámbito sigue estando abierto a la investigación, posiblemente debido, entre otros motivos, a que la gastronomía es tan popular como desafiante, donde las preferencias de las personas son muy variadas y están sesgadas por variedad de motivos, como son las propias preferencias, cuestiones culturales o salud. De esta manera se proponen diferentes estrategias en torno a la sugerencia de comida y recetas para cocinar. El artículo previamente mencionado, en el que se realiza una exploración de los métodos de recomendación más destacados [43], ha permitido expandir el proceso de revisión, aportando nuevos documentos que, junto con los ya recopilados en la sección anterior, nos permite elaborar la lista de artículos siguiente, los cuales resumen los métodos más populares e influyentes en la recomendación de comida hasta el momento.

En *Recipe Recommendation Using Ingredient Networks* [1], uno de los artículos más populares en sistemas de recomendación de comida, los autores hacen uso de la teoría de grafos y análisis de redes sociales para mejorar la precisión de las sugerencias de recetas de cocina. Como fuente de datos utilizan recetas extraídas del sitio web Allrecipes y las redes que construyen son las siguientes:

- Grafo de complementos de ingredientes, donde dos vértices -ingredientes- están conectados si han participado en la misma receta y el peso asignado a cada rama es proporcional a la coocurrencia de ambos ingredientes en todas las recetas de la colección.
- Grafo de sustituciones de ingredientes, construido a partir de aquellos comentarios de los usuarios donde se sugieren cambios en alguno de los ingredientes de la receta.

La investigación tiene como objetivo determinar entre parejas de recetas similares cuál de ellas tiene mayor valoración media y, por tanto, cuál de ellas es mejor candidata para ser recomendada. Para ello construyen diferentes versiones de un sistema de predicción, cuyo poder de predicción varía desde el 60% hasta 80%, siendo las versiones que mayor precisión consiguen las que utilizan el conocimiento obtenido a partir de los grafos mencionados. Con esto demuestran el poder predictivo que tienen las relaciones entre los ingredientes y el conocimiento aportado por las personas en torno a las sugerencias de reemplazo de unos ingredientes por otros.

Por su parte, en *User's Food Preference Extraction for Personalized Cooking Recipe Recommendation* [9], se presenta un método de extracción de las preferencias de usuario a través del estudio del uso de los ingredientes. Para este propósito formulan un conjunto de heurísticas que permiten determinar de forma numérica el grado de preferencia que tiene un usuario sobre cada ingrediente, expresada mediante las siguientes dos variables:

- I_k^- : grado en que al usuario no le gusta el ingrediente k.
- I_k^+ : grado en que al usuario le gusta el ingrediente k.

En el cálculo incorporan un modelo similar a TF-IDF, que combina la frecuencia de uso de cada ingrediente en los días más recientes y su especificidad en la colección. Asimismo proponen una métrica para predecir la puntuación de un usuario a una receta, que considera las variables anteriores y la similitud de la receta objetivo con todas las recetas que dicho usuario ha consumido en los últimos días, aportando como novedad el hecho de que repetir recetas en un corto período resulta determinante en la elección de una determinada receta.

El siguiente artículo, *Learning user tastes: a first step to generating healthy meal plans?* [15], consta de dos partes, una de ellas más orientada a la investigación y otra a la experimentación y evaluación. La primera parte está dedicada a la búsqueda y análisis de factores que puedan resultar determinantes a la hora de proporcionar un rating a una receta. Para ello los autores preparan un escenario de encuestas donde los encuestados deben valorar una serie de recetas, con el fin de averiguar cuáles son las distintas razones por las que los usuarios han proporcionado cada rating a las diferentes recetas que les han sido presentadas. Con este primer estudio consiguen una lista de 23 factores que han resultado influyentes en los usuarios, tanto de manera negativa como positiva, a la hora de valorar una receta, entre los cuales se encuentran como más relevantes los ingredientes que incluye la receta y tiempo de preparación y cocinado.

En la segunda parte de la investigación los autores tratan construir un algoritmo de recomendación de recetas que tenga en cuenta los factores más importantes de los 23 previamente identificados. Tal algoritmo es un sistema de recomendación basado en contenido y, utilizando combinaciones de matrices de usuarios, recetas y características -ingredientes-, consiguen una versión del algoritmo con un error MAE de 1,072 y RMSE de 1,256, mejorando así los resultados obtenidos con otros recomendadores, como aquellos construidos por los mismos autores basados en filtrado colaborativo y versiones más simples de algoritmos basado en contenido. Entre los factores más decisivos en la obtención de tal sistema están la incorporación de la información nutricional al proceso de predicción y la incorporación de los ingredientes como factor negativo, debido a que, según las encuestas que realizaron, los ingredientes influyen tanto positiva como negativamente a la hora de valorar una receta.

Otro sistema de recomendación basado en contenido es el diseñado por Freyne et al. [18] en *Recommending Food: Reasoning on Recipes and Ingredients.* Este sistema crea un perfil de usuario compuesto de un vector con tantas dimensiones como ingredientes distintos existan en la colección de datos. El valor asociado a cada dimensión constituye un valor promedio de los ratings de aquellas recetas en las que esté presente el ingrediente. Finalmente, cada uno de estos valores se promedian para predecir el rating de usuario a cada nueva receta objetivo.

De esta manera el algoritmo consigue mediante un método simple propagar las preferencias de usuario hacia los ingredientes, y desde éstos hacia los nuevos ítems a recomendar.

En el artículo además se toma el algoritmo anterior como base para crear otros dos algoritmos híbridos, que aun combinando la técnica de filtrado colaborativo con el proceso de propagación de ratings anteriormente mencionado, no consiguen mejorar la métrica MAE de 0.262 del algoritmo base.

Este experimento demuestra que partiendo de una idea intuitiva se puede obtener un sistema de recomendación de recetas eficaz. Sin embargo, este método no tiene en cuenta otros factores también importantes en el gusto por los ingredientes, como pueden ser la ausencia o presencia de un ingrediente en una receta o la frecuencia de su uso, que pueden generar un mayor conocimiento en torno a las preferencias de las personas por los ingredientes y, en general, por la comida.

En el artículo *Content boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation*, se hace uso de la técnica de factorización de matrices en el ámbito de recomendación de recetas de cocina [22]. Inspirados por la competición *The Netflix Prize* [44] y la solución ganadora [45], cuyo algoritmo utilizaba esta técnica, Forbes et al. han trasladado esta metodología al mundo gastronómico y culinario. Para ello utilizan los usuarios, los ingredientes, los ratings y las propias recetas (ítems) como únicas entidades presentes en los datos.

En el proceso construyen a partir de los datos la matriz de ratings (usuarios x recetas), y constituye la matriz que se desea factorizar de la siguiente manera:

$$S \approx UR^T,$$

siendo *U* la matriz de características de los usuarios y *R* la matriz de características de las recetas. A su vez y con el fin de explotar la información que se dispone de los ingredientes, descomponen *R* en el producto de otras 2 matrices *X* y Φ que constituyen la presencia o ausencia de un ingrediente en la receta y el vector de características de cada ingrediente, respectivamente. Utilizando la técnica de descenso por gradiente para estimar los valores de las matrices U y Φ, consiguen mejorar en RMSE una versión base del algoritmo de filtrado colaborativo, además de capturar relaciones de similitud y combinación entre los ingredientes que van más allá de la medición de la coocurrencia en las distintas recetas.

Como se ha podido observar, a partir de las anteriores investigaciones, los ingredientes de las recetas constituyen una fuente de información con alto potencial, capaz de proporcionar información extensa de las preferencias de las personas. Sin embargo los datos de una receta contienen más información que puede resultar útil a la hora de sugerir nuevas recetas de cocina, tal y como demuestran El-Dosuky et al. en *Food Recommendation using Ontology and Heuristics* [24], en cuyo experimento construyen diferentes versiones de un sistema de recomendación semántico de recetas de comida. En esta propuesta se toma de partida una ontología de gastronomía y una serie de heurísticas nutricionales, que reflejan reglas que deben seguirse hacia un hábito de comida saludable. El sistema toma la descripción de cada una de las recetas con las que ha interactuado cada usuario y realiza un preprocesamiento con el fin de obtener las palabras relevantes de cada una.

En este punto diferentes modelos de procesamiento de palabras son utilizados, como son TF-IDF, Jaccard y coseno binario, así como un modelo propuesto basado en teoría de conjuntos entre las palabras de la receta, las del usuario y las heurísticas mencionadas.

La evaluación de los distintos modelos demuestra finalmente que la versión que los mismos autores proponen presenta mejores resultados, con un 94% *accuracy* y 94% de precisión. La evaluación es online realizada con 5 personas y demuestra que la descripción de las recetas de comida también permite a un sistema generalizar conocimiento y predecir el gusto en la comida.

Los métodos de factorización de matrices han tenido un gran impacto en torno a la minería de datos y los sistemas de recomendación. En el artículo *Using Tags and Latent Factors in a Food Recommender System* [46] se propone, en primer lugar, el diseño de interacción de una interfaz de usuario de una aplicación para dispositivos móviles, capaz de recoger preferencias de usuario en recetas de cocina en forma de ratings y etiquetas. Estas preferencias constituyen los datos de entrada de un sistema de recomendación de recetas de comida que utiliza como modelo base la factorización de matrices de ratings de usuarios a recetas.

En la interfaz de usuario diseñada, el usuario es capaz de proporcionar un rating numérico a una receta y además añadir un conjunto de etiquetas, pudiendo contener cualquier término, junto con una connotación (positiva o negativa) para cada etiqueta. De esta manera se puede obtener tanto las etiquetas asignadas por los usuarios a cada receta como su polaridad.

Por otro lado, el sistema de recomendación está basado en el modelo descrito por Ignacio e Iván de filtrado colaborativo aplicable a múltiples dominios incorporando asignación de etiquetas [47]. De esta manera los autores consiguen introducir ambos conjuntos de etiquetas (etiquetas por usuario y etiquetas por receta) en el cálculo base del rating, consiguiendo un sistema con un error MAE promedio de 0,686, mejorando el obtenido mediante la versión más básica del modelo de factorización de matrices con 1,018.

Una vez más se puede observar el potencial de los métodos de factorización de matrices aplicados en sistemas de recomendación de comida, incluyendo tan sólo el uso de etiquetas y su polaridad como factores latentes del modelo, los cuales permiten realizar predicciones aunque el usuario no haya asignado etiquetas a ninguna receta del sistema. Sin embargo este algoritmo es susceptible de mejora ya que esta versión sólo utiliza las etiquetas con polaridad positiva, donde las etiquetas negativas pueden ampliar la capacidad de generalizar conocimiento acerca de los gustos de las personas.

Estado del Arte

A partir de la revisión realizada se pueden destacar las siguientes conclusiones:

- *Diversidad de modelos.* Para hacer frente al problema de la sugerencia de recetas en el contexto gastronómico, las diferentes investigaciones han recurrido a variados modelos de recomendación, tanto basados en contenido, utilizando desde grafos de ingredientes hasta propagación de ratings, como métodos de filtrado colaborativo e híbridos, que combinan el uso técnicas de factorización de matrices con datos de las recetas para capturar las preferencias del usuario.

- *Ingredientes.* Los ingredientes constituyen el componente más importante de una receta de cocina y por tanto donde las personas reflejan con mayor fuerza sus preferencias en comida. Asimismo resulta ser el núcleo de información para sistemas de recomendación basados en contenido, y así lo demuestran estos trabajos, donde en todos se incorpora y se modeliza el uso de los ingredientes para sugerir nuevos elementos.

- *Salud.* En algunos de los experimentos se ha intentado introducir en los algoritmos de recomendación información sobre salubridad, en su mayoría bajo la forma de datos nutricionales. La salubridad de la comida es un terreno que con el tiempo está tomando mayor relevancia, lo cual justifica que en los nuevos avances en sistemas de recomendación de recetas se dedique esfuerzo en incorporar la diferenciación entre recetas saludables y no saludables.

- *Datos*. En la gran mayoría de los artículos los autores han realizado procesos propios de obtención de datos, lo cual confirma la premisa formulada al inicio de este trabajo tras una breve revisión de los artículos más destacados.

- *Alcance*. Utilizando como dato base los ingredientes, se elaboran algoritmos con una precisión más que aceptable. Sin embargo el alcance de la minería de datos en gastronomía es más amplio de lo que han abarcado estos experimentos, si bien los gustos de las personas tienen en cuenta variables más complejas, como pueden ser patrones temporales, cultura, ubicación geográfica, sentimientos, moda o la influencia social.

- *Historial de usuario*. De entre los artículos revisados que tratan temas relacionados con sistemas de recomendación, casi la totalidad utilizan datos que carecen de información sobre el historial del usuario, entendiendo el historial de usuario como un registro de su actividad. A modo de ejemplo, el historial de usuario en aplicaciones de compartición de recetas incluiría el listado de recetas que dicho usuario ha compartido. Estos datos pueden resultar útiles en sistemas de recomendación basados en contenido, ya que pueden permitir incorporar información sobre ítems que, aunque el usuario no le haya asignado una valoración, podemos decir que lo ha consumido.

3. Inferencia de dificultad

La dificultad de una receta resulta relevante a la hora de elegir hacerla, y es por ello que constituye otra información susceptible de tener en cuenta en un sistema de recomendación. Es una información implícita que, nutriéndose de datos como los pasos de una receta (en número y duración), la cantidad de ingredientes utilizados y la complejidad que supone disponer de ellos, los instrumentos de cocina utilizados o incluso los propios métodos de cocinar implicados, ofrece distintas posibilidades a la hora de incorporar las preferencias de usuario y de agregar información de distintos tipos de datos. Es por ello por lo que en la comunidad científica también se ha dedicado esfuerzo en esta dirección y en el proceso de revisión del estado del arte se ha encontrado el artículo *Complexity and Similarity of Recipes based on Entropy Measurement.*

En este artículo [48] se utiliza la teoría de la información para inferir tanto la dificultad de una receta como para medir la similitud entre dos recetas. Tras realizar un análisis estadístico de los ingredientes y los verbos de cocina utilizados en los datos de recetas, los autores utilizan los ingredientes y verbos como variables aleatorias, lo cual les permite calcular su entropía, que no es sino una medida de la cantidad de desinformación de que se dispone sobre cierta variable aleatoria. La entropía medirá en ambos casos cómo de raro es el objeto de estudio, para lo cual la entropía de una receta se puede calcular agregando la entropía de los subyacentes verbos de cocina y los ingredientes que incluye. La entropía de una receta por tanto resulta ser una medida de cómo de comunes o raros son sus componentes principales; de este modo una receta con una entropía alta incluye ingredientes poco comunes y métodos de preparación y cocinado menos frecuentes y, por ello, posiblemente resulte ser más compleja que otra receta con menor entropía.

Este método también utiliza la entropía *E* de los ingredientes de las recetas para calcular la similitud con otras mediante la siguiente expresión:

$$sim(r_x, r_y) = \frac{\sum_{i_k \in \{r_x \cap r_y\}} E(i_k)}{\sum_{i_k \in \{r_x \cup r_y\}} E(i_k)}$$

, donde i_k representa cada ingrediente de las recetas r_x y r_y cuya similitud se pretende medir. A pesar de no realizar un proceso de evaluación, los autores calculan la entropía de las recetas de su dataset tanto de los verbos como de los ingredientes utilizados, obteniendo recetas como “Lasaña” o “Lasaña vegetal” con valores muy altos de entropía y “Patata cocida” o “Panettone” con valores muy bajos. Con esto se demuestra que un método simple de Teoría de la Información resulta ser una heurística aceptable para la inferencia de dificultad en las recetas de cocina, teniendo en cuenta los dos componentes principales que influyen en esta propiedad, como son los ingredientes y los métodos de cocinado.

Capítulo 3

Definición del conjunto de datos

En vista de la dificultad para obtener un conjunto de datos a partir de fuentes públicas de recetas de cocina, el primer paso ha consistido en la definición del conjunto de datos que deseamos obtener, explorando las distintas fuentes de datos disponibles, de acuerdo con el objetivo 2 detallado más arriba. Ambos procesos, definición y exploración, han sido realizados al mismo tiempo de acuerdo con la lista de restricciones, que veremos a continuación, determinada tras comprobar el alcance de cada una de las fuentes de datos visitadas.

1. Restricciones

Las fuentes de datos exploradas han sido en su totalidad sitios web, que abarcan desde blogs de cocina hasta redes sociales, algunos de los cuales se han empleado para la obtención de datos ad hoc de los trabajos que se han revisado anteriormente. Las restricciones que se han impuesto para la extracción de los datos son tanto de carácter obligatorio como opcional, enfocadas a la obtención de un conjunto de datos (dataset) completo que pueda ser utilizado tanto en este trabajo como en futuros experimentos.

Contenido imprescindible

La siguiente lista recoge los distintos elementos que deben contemplar las fuentes de datos, y que constituyen la información básica de las recetas y usuarios para poder aplicar algoritmos básicos de filtrado colaborativo y basados en contenido.

- **Recetas**. Las entidades principales del conjunto de datos serán las recetas de cocina compuestas por, al menos, los siguientes atributos:
    - Título de la receta.
    - Pasos o instrucciones de la receta. La receta debe tener instrucciones claramente identificadas y separadas.
    - Ingredientes utilizados en la receta, cada ingrediente por separado, junto con la cantidad asociada.
- **Existencia de usuarios**. El conocimiento que se desea generalizar, tanto a nivel de recomendaciones como a nivel de análisis de datos en general, está orientado hacia los usuarios (modelos de consumo de comida, salud, sugerencia de comida, etc.). Lo usuarios deben ser quienes publiquen las recetas y puedan interactuar con ellas.
- **Perfil de usuario**, con cierta información personal (geolocalización, lista de recetas, lista de revisiones).
- **Valoración promedio** de una receta.

- **Valoración numérica** asignada a una receta, con un rango entre 1 y 5, entendiendo por valoración como una calificación que un usuario asigna a una receta.
- **Revisiones** o comentarios que los usuarios realizan a las recetas.
- **Datos nutricionales**, con al menos la información nutricional básica de cada receta, incluyendo número de calorías y cantidad de hidratos de carbono, grasas y proteínas.

Contenido deseable

A continuación se detallan los diferentes atributos que se desea incorporar en el conjunto de datos, los cuales incrementarían su utilidad, permitiendo escenarios de experimentación más complejos.

Receta

- **Descripción de la receta**, donde el usuario puede proporcionar cualquier tipo de información acerca de la receta, como puede ser un resumen, ingredientes y métodos principales, contexto social donde mejor se adapta, o incluso su opinión.
- **Tiempo de preparación y cocinado.**
- **Número de valoraciones** que ha recibido la receta.
- **Número de valoraciones** que ha recibido la receta por cada valor del rango numérico.
- **Consejos**. Lista de consejos, notas o posibles variantes a tener en cuenta por aquellos otros usuarios que deseen incorporarla en su menú.
- **Etiquetas** que el autor asigna a la receta, que pueden tener cualquier contenido.
- ***Me gusta*/*No me gusta***. Valoración binaria a través de la cual los usuarios declaran si le gusta o no le gusta la receta.
- **Número de raciones**.
- **Tipo de plato** (principal, primer plato, postre, etc.).
- **Ocasión**. Ocasión donde mejor se adapta la receta (Navidad, comida social, etc.).
- **Información nutricional extendida**, incluyendo una lista de nutrientes y su cantidad.
- **Categoría**, siendo válido cualquier tipo de categorización (cultura, país, tipo de plato, ingrediente principal, método, etc.).
- ***La he hecho***. Número de usuarios que declararon hacer la receta.
- **Recetas relacionadas**. Recetas que la aplicación considera que están relacionadas con la receta actual.
- **Salubridad**. Un tipo de categoría de recetas saludables, o bien una propiedad adicional inherente a las recetas que indique si es considerada o no saludable.
- **Dificultad de la receta**, en forma numérica o categórica, de al menos 3 valores.

Revisión

- **Revisión con valoración**. Posibilidad de que al realizar una revisión se proporcione una valoración numérica o rating.
- **Valoración sin revisión**. Posibilidad de valorar la receta si necesidad de escribir una revisión.
- ***Me gusta*/*No me gusta***. Valoración binaria a una revisión, que indique su grado de relevancia o utilidad para el resto de usuarios.

Definición del conjunto de datos

Usuario

- **Lista de recetas favoritas**. Recetas que el usuario declaró que le gustaron, ya sean propias o ajenas.
- **Revisiones/valoraciones realizadas**. Lista de revisiones y valoraciones realizadas por el usuario.
- **Estructura de red social**. Los usuarios pueden tener relación de interés culinario o amistad con otros usuarios. Esta característica resulta de interés desde el punto de vista de la Minería de Datos, ya que ofrece la posibilidad de aplicar diversos algoritmos orientados a redes sociales.
- **Valoración media**. Valoración promedio de todas las valoraciones que han tenido las recetas de un usuario.
- **Perfil público**. Posibilidad de que el perfil del usuario sea totalmente público, con el fin de facilitar el proceso de extracción de datos.

Una vez elaborada esta lista de requerimientos se ha realizado una exploración de los sitios web más populares, analizando en qué medida los cumplen. Los sitios web explorados son los siguientes:

- Allrecipes[7]
- Foodnetwork[8]
- Food.com[9], fusionada actualmente con otra aplicación en Genius Kitchen.
- Yummly[10]
- Betty Crocker[11]
- Kraft Recipes[12]
- MyRecipes[13]
- EatingWell[14]
- Cookpad[15]
- Cooks[16]
- Simply recipes[17]

Para cada fuente de datos se ha determinado si cumple o no cada tipo de requisito. Hemos asignado puntuaciones de acuerdo con las graduaciones siguientes: asignamos 3 puntos al contenido imprescindible, 2 al contenido deseable y 1 al contenido prescindible. El anexo A contiene las tablas que se han construido para decidir qué fuente de datos se va a utilizar, y reúnen un conjunto total de todas las características que han sido observadas en todas las fuentes de datos.

[7] https://www.allrecipes.com

[8] https://www.foodnetwork.com

[9] http://www.geniuskitchen.com/

[10] https://www.yummly.com/

[11] https://www.bettycrocker.com/

[12] http://www.kraftrecipes.com/

[13] http://www.myrecipes.com/

[14] http://www.eatingwell.com/

[15] https://cookpad.com/es

[16] http://www.cooks.com/

[17] https://www.simplyrecipes.com/

Según el criterio de puntuación utilizado la fuente de datos elegida ha sido Allrecipes que, además, constituye la fuente de datos más utilizada en los artículos revisados, según se destaca en *Food Recommender Systems* [43].

2. Allrecipes

Allrecipes es una red social vertical centrada en el ámbito culinario. Es uno de los portales web de cocina más visitados en el mundo, con aproximadamente 25 millones de visitantes mensuales[18], operando en 19 sitios web en 13 idiomas distintos[19].

En un primer proceso de revisión hemos comprobado que Allrecipes dispone de una API REST[20] (requiere credenciales para su uso) que permite el acceso a los recursos de la web.

Una vez determinado el portal del cual realizar la extracción de datos se ha realizado un estudio de la viabilidad técnica de la extracción, ya que en la actualidad muchos sitios web generan su contenido de manera dinámica con lenguajes de programación como Javascript, lo cual dificulta el proceso de extracción. En este proceso se han generado las tablas 2 y 3 que resumen la relevancia y disponibilidad técnica de los principales contenidos, entendiendo por relevancia la importancia que se ha asignado a cada contenido.

Contenido	Disponibilidad	Relevancia
Lista de recetas	HTML	Alta
Nombre	HTML	Alta
Valoración media	HTML	Alta
Descripción	HTML	Alta
Ingredientes	HTML	Alta
Tiempos	HTML	Alta
Información nutricional básica	HTML y JS	Alta
Pasos	HTML	Alta
Revisiones	JS	Alta

Tabla 2: Disponibilidad técnica de la extracción del contenido de una **receta** en Allrecipes

Contenido	Disponibilidad	Relevancia
Lista de recetas	HTML	Alta
Nombre	HTML	Alta
Número de seguidores	HTML	Media
Lista de seguidores	HTML	Alta
Lista de revisiones	JS	Alta
Biografía	HTML	Alta

Tabla 3: Disponibilidad técnica de la extracción del contenido de un **usuario** en Allrecipes

[18] http://www.ebizmba.com/articles/recipe-websites

[19] https://en.wikipedia.org/wiki/Allrecipes.com

[20] https://apps.allrecipes.com/

Definición del conjunto de datos

La mayor parte del contenido de Allrecipes, a excepción de las listas de revisiones y la información nutricional detallada, está reflejado en el código HTML que el servidor envía como respuesta, lo cual ha constituido razón suficiente para elegir definitivamente Allrecipes como fuente de datos objetivo.

Además en este proceso se ha explorado el sitio web en su totalidad, observando todas las funcionalidades de interés que dispone al usuario dentro del portal. El sitio web tiene una amplia gama de categorías de recetas, accesibles y filtrables vía HTML y URL, lo cual resulta interesante para poder extraer recetas ya categorizadas.

Definición del conjunto de datos

Capítulo 4

Software de extracción de datos

En este capítulo se detalla el proceso de extracción de datos que se ha realizado para obtener el dataset deseado. Una vez elegida como fuente de datos objetivo Allrecipes, se ha procedido a realizar la extracción de datos masiva. Para ello se ha construido un software automático que realiza las labores de *crawling* y *scraping*, con el fin de obtener todos los datos que vamos a requerir. A continuación se describen las fases de desarrollo del software, así como el detalle de la ejecución y de los resultados obtenidos.

1. Análisis

Allrecipes es un sitio web de recetas de comida con estructura de red social. Los usuarios publican recetas y pueden interactuar con las recetas que publican otros usuarios para asignarles una valoración, realizar un comentario o crítica en forma de revisión, o incluso declarar que les ha gustado o han cocinado la receta en cuestión. Los usuarios a su vez tienen un perfil compuesto por los usuarios que siguen su actividad (seguidores) y aquellas personas por las que muestran interés (amigos), y es lo que proporciona al sistema la estructura de red social. Además el sitio web dispone de una jerarquía de categorías a través de la cual el usuario puede buscar y filtrar recetas, por lo que cada receta tiene asignada una categoría. Se ha decidido incorporar la jerarquía de categorías de recetas al dataset porque contiene una agrupación de recetas que puede resultar útil para posteriores procesos analíticos. De este modo en el proceso de extracción se desea capturar la información que contiene el sistema en torno a estas cuatro entidades: recetas, revisiones, usuarios y categorías.

1.1 Estructura de la información

El primer lugar, mediante un proceso de ingeniería inversa, hemos capturado la manera en que están estructuradas las categorías, recetas y usuarios dentro de Allrecipes, así como las relaciones que existen entre ellas. Para tal cometido se ha realizado una exploración del sitio web, navegando entre las distintas páginas y observando los cambios en las URLs y los enlaces que aparecen en el código HTML generado en cada petición. En este proceso también se ha detectado que las diferentes entidades que se van a extraer (categorías, recetas, usuarios y revisiones) disponen de identificadores, aparentemente únicos, generados por el propio sistema *backend* de Allrecipes, lo cual facilita el proceso de extracción de los datos. A continuación se describe la estructura encontrada para cada una de las entidades mencionadas.

Categorías

En Allrecipes existe una jerarquía de categorías de recetas, donde cada receta pertenece a una sola categoría. A su vez, una categoría puede pertenecer (o tener como categoría padre) a una o más categorías, y es lo que otorga a esta jerarquía la estructura de grafo, tal y como puede observarse en la Ilustración 1. La categoría se le asigna a cada receta en el momento de ser publicada por un usuario.

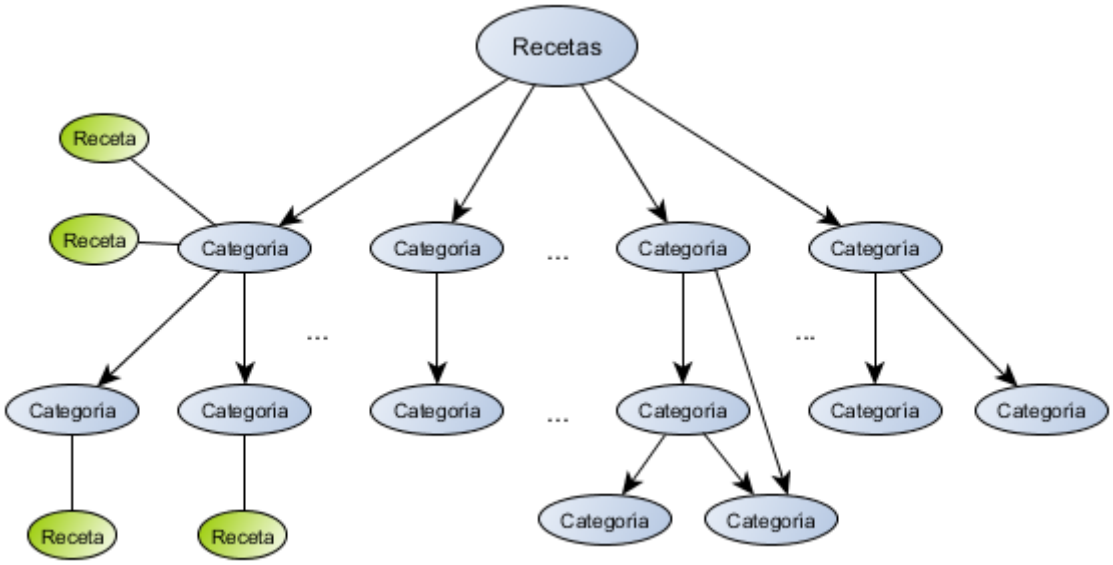

Ilustración 1: Estructura de las categorías en Allrecipes

Receta

En la Ilustración 2 se detalla la estructura básica de una receta en Allrecipes, donde podemos ver que, además de su información básica, tiene asociadas una lista de revisiones y una lista de recetas similares. Los detalles de los distintos atributos que serán extraídos serán descritos en la fase de diseño.

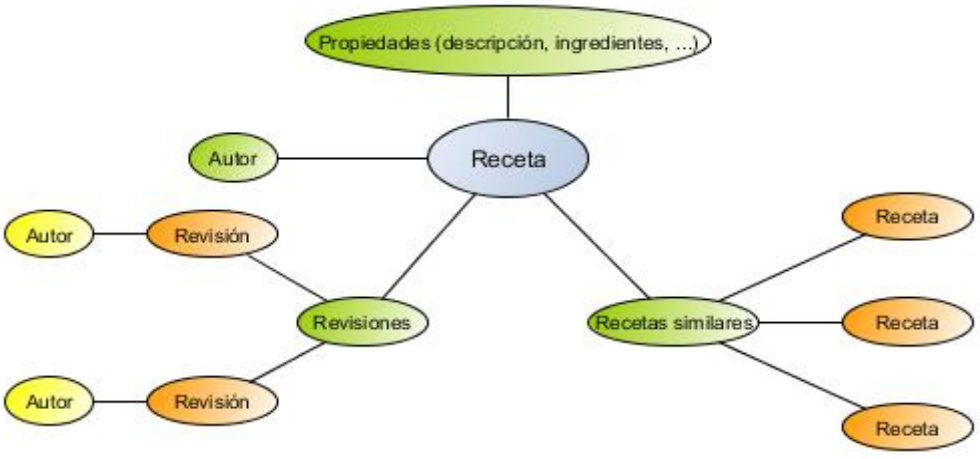

Ilustración 2: Estructura de una receta en Allrecipes

Usuario

Ilustración 3: Estructura de un usuario en Allrecipes

La Ilustración 3 muestra la estructura de un usuario en Allrecipes. Como se puede ver, el usuario, además de las características básicas (detalladas en la fase de diseño), tiene una lista de usuarios 'seguidores', que son usuarios que declaran seguir su actividad, y lista de usuarios 'amigos', que son aquellos por los que el usuario muestra interés.

Además el perfil de usuario dispone de tres listas de recetas (recetas que le gustaron, recetas que declaró cocinar y recetas publicadas) así como de una lista de revisiones que ha emitido sobre otras recetas.

Interacción Usuario-Usuario

La interacción entre usuarios, tal y como puede deducirse a partir de la estructura de un usuario, consiste en una relación de seguimiento o amistad, de acuerdo con la siguiente Ilustración 4:

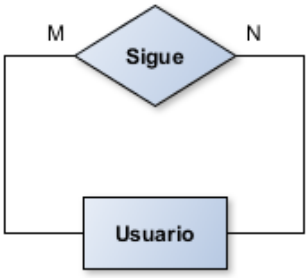

Ilustración 4: Relación entre usuario-usuario en Allrecipes

Interacción Usuario-Receta

Por otro lado, los usuarios pueden interactuar con las recetas de diversas maneras: creación (y publicación) de una receta, revisión de una receta, marcar como favorita (‘me gusta’) una receta y declarar (*madeit)* que ha realizado una receta. La cardinalidad de estas relaciones son M-N a excepción de la de ‘Publica’, donde una receta solo puede ser publicada por un mismo usuario. El diagrama de la Ilustración 5 ilustra las relaciones mencionadas:

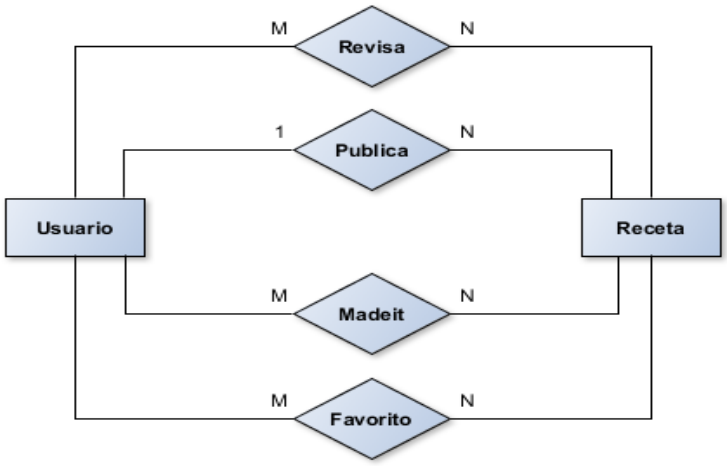

Ilustración 5: Relación entre usuario y receta en Allrecipes

Diagrama Entidad-Relación

Una vez definidas la estructura y las relaciones de las entidades que se van a extraer (Categoría, Usuario, Receta y Revisión), se ha deducido el siguiente diagrama Entidad-Relación, que recoge de manera general todas estas relaciones (Ilustración 6). En este diagrama se omiten los atributos de las entidades ya que serán descritos en detalle en el apartado de diseño.

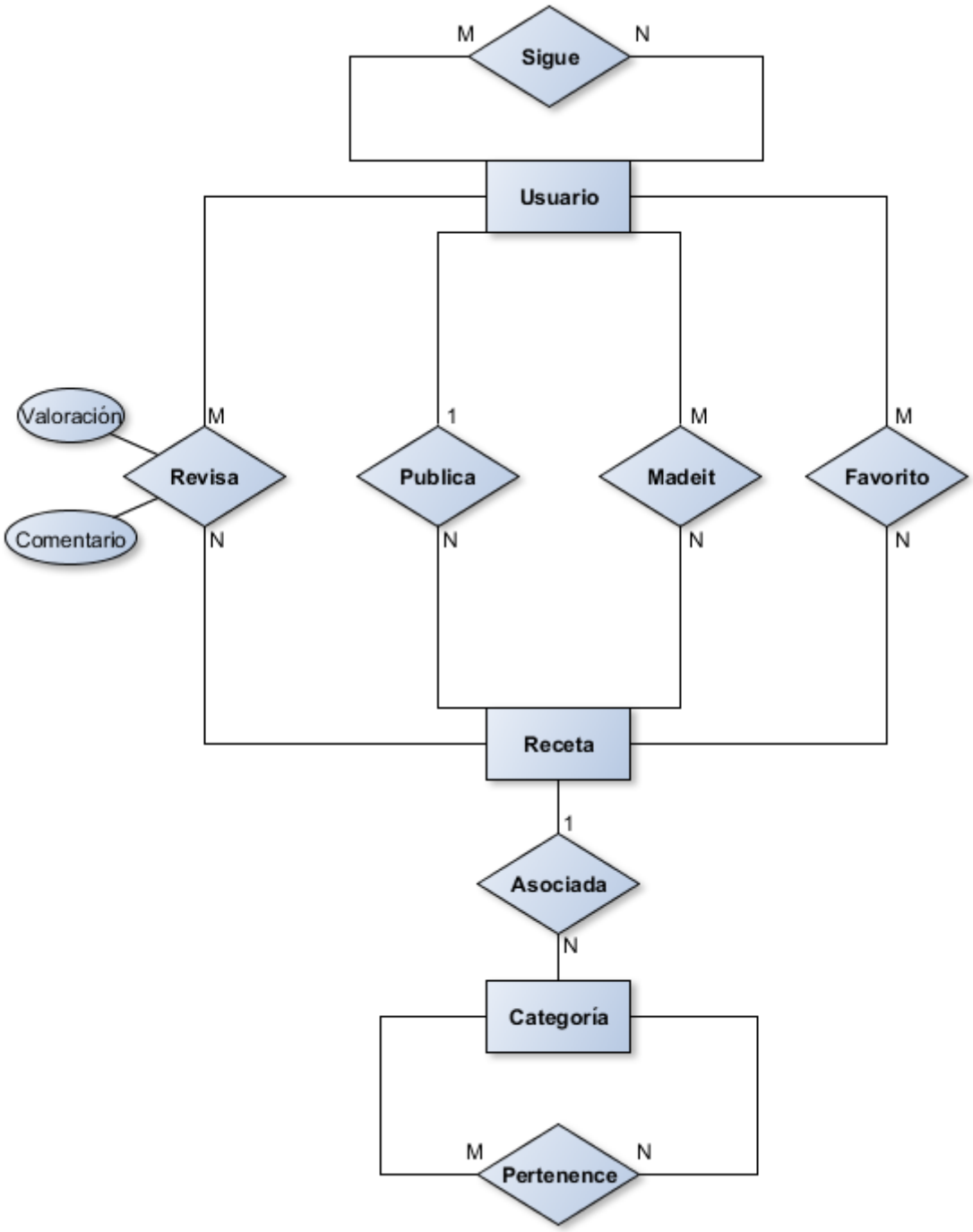

Ilustración 6: Diagrama Entidad-Relación deducido de Allrecipes

En el diagrama se ha incluido la entidad *Categoría*, que mediante la relación *Pertenece* permite reflejar la jerarquía en forma de grafo de las categorías de Allrecipes. Esta entidad además está relacionada con la entidad *Receta*, expresando que una receta tiene asignada una categoría y que a una categoría pueden estar asociadas varias recetas. Los únicos atributos que se han incluido son aquellos asociados a la relación *Revisa*, donde en el momento en que un usuario escribe una revisión a una receta, debe incluir un texto o comentario y una valoración numérica asociados a la revisión.

1.2 Contenido generado dinámicamente

Tal y como se ha observado en el análisis de viabilidad y disponibilidad de la información de Allrecipes (capítulo 3), existe cierta información en el portal que se genera dinámicamente y que no está disponible en el código HTML de las páginas web. Esta información incluye datos como la información nutricional de cada receta de manera extendida y la lista de revisiones de cada receta, entre otros. Ante la imposibilidad de obtener estos datos de manera rápida vía HTML, se ha realizado otro proceso de ingeniería inversa, con el fin de poder disponer de esta información de otra manera.

El contenido dinámico de las páginas web Allrecipes se genera en el lado del cliente (navegador web) a partir de ficheros en formato JSON que el cliente web solicita para cargar la página. El primer paso ha sido examinar en detalle las peticiones realizadas solicitando dicho contenido, incluyendo las diferentes URLs y sus parámetros, y el contenido de los ficheros. En este primer paso se ha visto que el dominio de las URLs en este tipo de peticiones es diferente al dominio utilizado para cargar las páginas web, siendo el primero de ellos *allrecipes.com* y el último *apps.allrecipes.com*. Se ha deducido que el último de estos dominios corresponde a la API de Allrecipes mencionada anteriormente, y constituye una API REST a la cual Allrecipes solicita algunos de los datos necesarios para cargar las distintas páginas web.

Seguidamente, mediante un cliente HTTP se ha inspeccionado qué datos están disponibles en la API y cuáles son las URLs asociadas. Las tablas que se muestran a continuación recogen el resultado de esta inspección, en las cuales se han omitido las URLs para evitar publicar datos protegidos.

Datos de usuario

Dato	Disponibilidad
Perfil (id, nombre, ciudad, país, etc.)	HTML y API
Lista de revisiones	API
Lista de publicaciones	API
Lista de favoritos	API
Lista de recetas cocinadas	API
Lista de seguidores	API
Lista de amigos	API

Tabla 4: Disponibilidad de los datos de usuario en Allrecipes teniendo en cuenta la inspección de la API

Como se puede observar, la única información que está disponible vía HTML es la información básica del usuario, que al mismo tiempo también está disponible a través de la API. En la Ilustración 7 se puede ver la página web donde se muestra el perfil de un usuario, donde se puede observar las tres listas principales de sus recetas (“Favorites”, “I Made It” y “Personal Recipes”), además de las revisiones, sus seguidores y sus amigos.

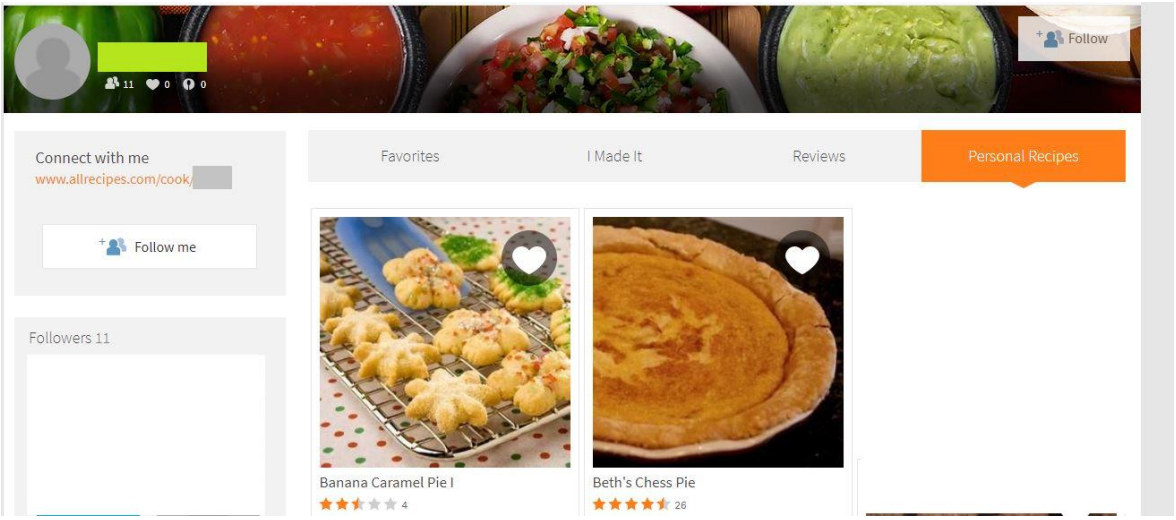

Ilustración 7: Ejemplo de la página web de perfil de usuario en Allrecipes

Teniendo en cuenta que no toda la información de usuario está disponible en la página web, se ha decidido extraer la información de usuario utilizando como único recurso de datos la API.

Datos de receta

Dato	Disponibilidad
Identificador	HTML y API
Título	HTML y API
Nombre e identificador del autor	HTML y API
Valoración media	HTML y API
Número de valoraciones	HTML y API
Número de revisiones	HTML y API
Número de personas que la cocinaron	HTML y API
Descripción	HTML y API
Número de raciones	HTML y API
Información nutricional básica (#calorías, nutriente y cantidad)	HTML y API
Información nutricional extendida (nutriente y cantidad)	API
Ingredientes (identificador y cantidad)	HTML y API
Tiempo de preparación	HTML y API
Tiempo de cocinado	HTML y API
Tiempo total	HTML y API
Pasos (orden y descripción)	HTML y API
Notas del autor	HTML y API
Identificador de categoría	HTML
Lista de revisiones	API
Identificadores de recetas similares	HTML y API

Tabla 5: Disponibilidad de los datos de receta en Allrecipes teniendo en cuenta la inspección de la API

Por su parte, la información de las recetas está disponible casi en su totalidad vía HTML, a excepción de la información nutricional extendida y la lista de revisiones (tal y como se vio en el capítulo anterior) y que, al mismo tiempo, están disponibles a través de la API, lo cual requiere hacer uso de esta API para obtener la información completa de las recetas. En las ilustraciones siguientes se muestra el contenido de una receta en Allrecipes, donde se puede ver la información básica de la receta, los pasos a seguir y los ingredientes.

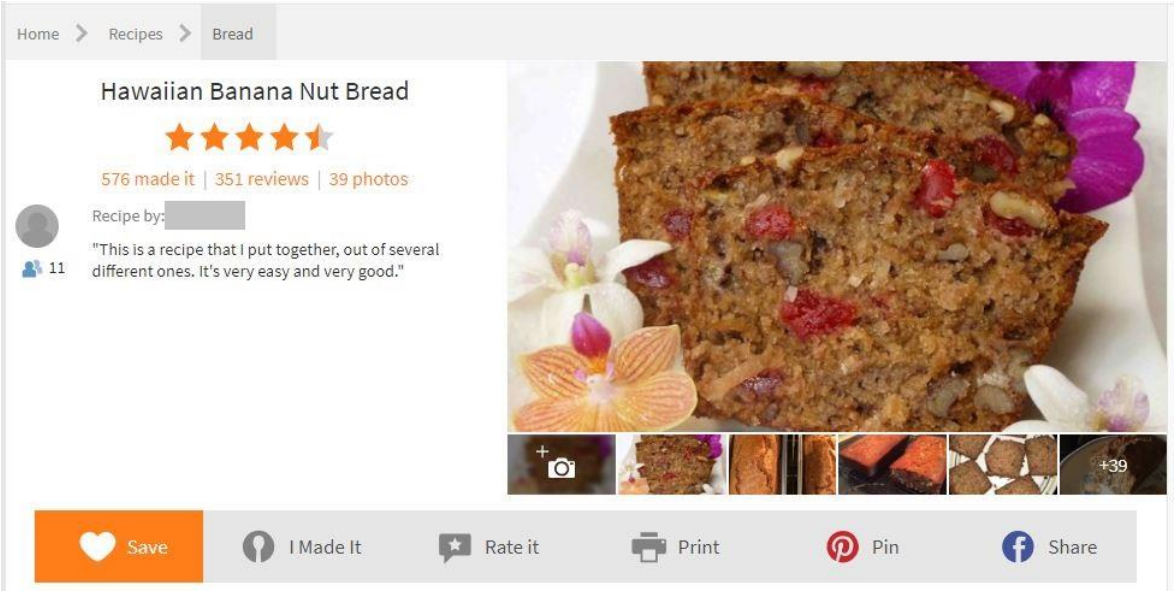

Ilustración 8: Ejemplo de receta en Allrecipes (1)

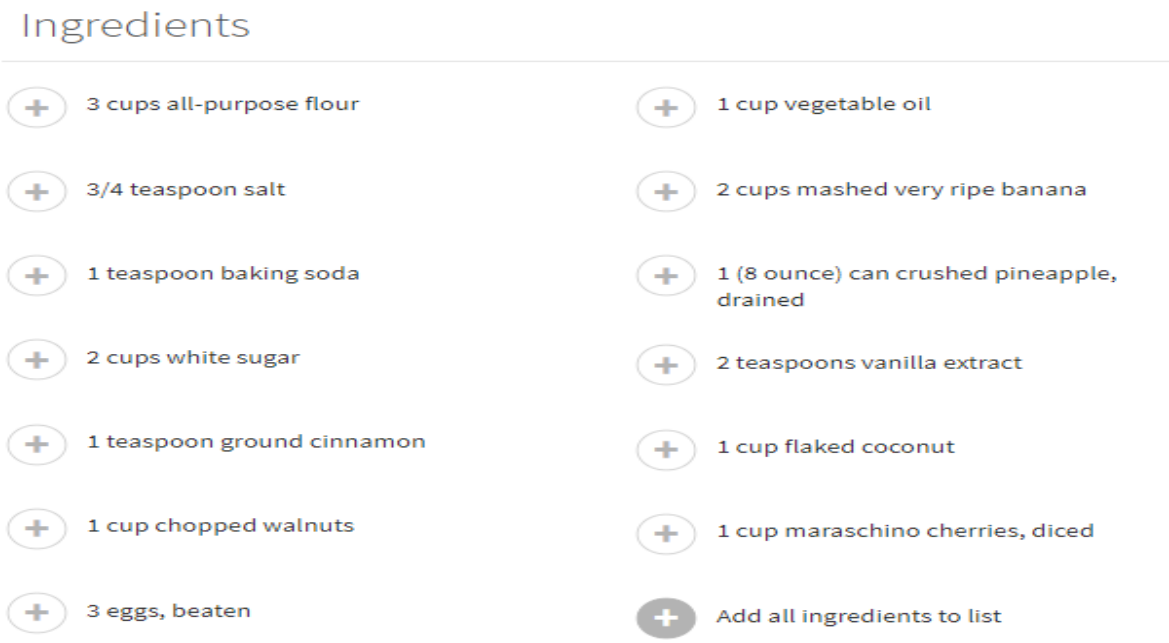

Ilustración 9: Ejemplo de receta en Allrecipes (2)

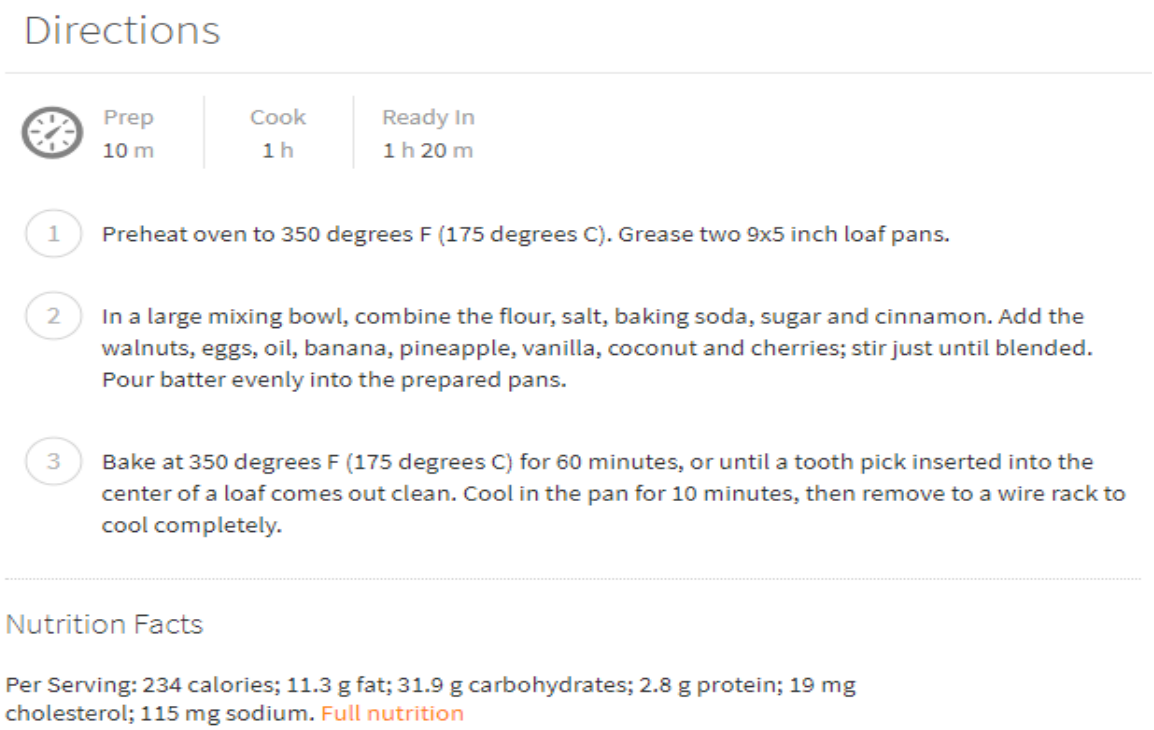

Ilustración 10: Ejemplo de receta en el sitio web de Allrecipes (3)

En el caso de la información asociada a la receta, y al contrario de como ocurre con la información de usuario, es necesario contar con la información contenida en el código HTML para poder extraer la categoría a la que pertenece la receta, lo que se traduce en la utilización de ambas vías (API y HTML) para capturar de manera completa la información de las recetas.

Datos de las categorías

En el proceso de revisión de la API no se ha encontrado ninguna URL disponible para solicitar información de las categorías de recetas. Esta información se encuentra disponible únicamente a través del código HTML de las páginas web de las distintas categorías. En el código HTML de cada una de estas páginas se muestra un contenido en formato JSON que contiene la lista de categorías, incluyendo URLs, nombres e identificadores numéricos, que son descendientes de la categoría visitada, dentro de la jerarquía de categorías. La Ilustración 11 muestra un ejemplo de categoría en la página principal de Allrecipes.

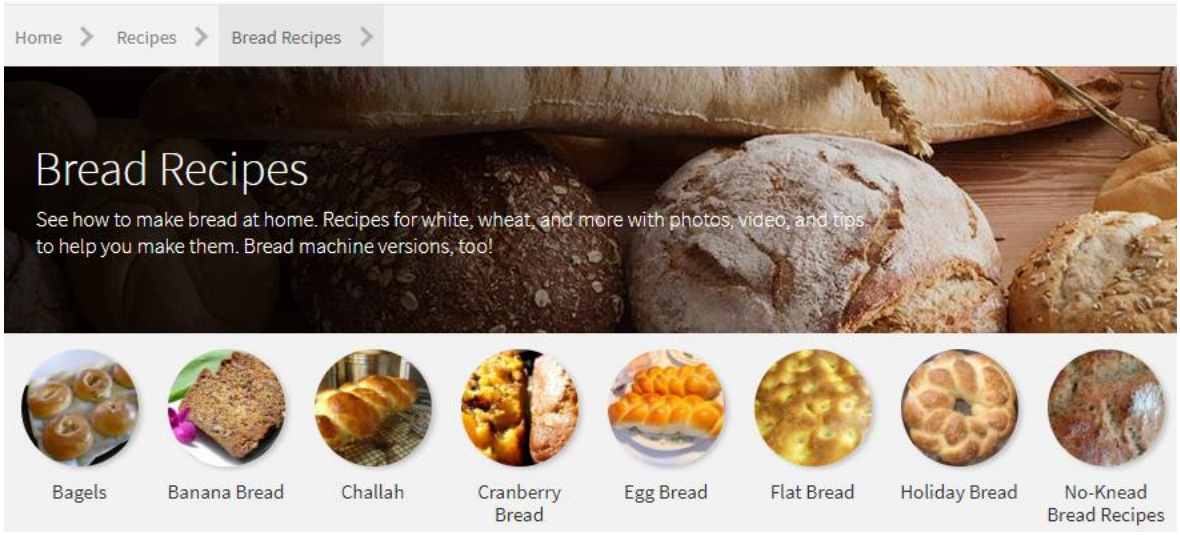

Ilustración 11: Ejemplo de jerarquía de categorías de la categoría “Bread Recipes” en Allrecipes

Debido a la inexistencia de datos a través de la API, y con el fin de disponer de datos sobre las categorías de recetas, resulta indispensable la extracción a partir del código HTML de las páginas web de las categorías de Allrecipes.

Autorización de la API

El último paso en el proceso de inspección de la API, una vez se ha establecido el catálogo de datos al que se puede acceder, ha sido examinar los distintos requisitos de autorización, acceso y utilización subyacentes. La API de Allrecipes implementa el protocolo OAuth2[21], que constituye un protocolo de autorización para habilitar el acceso limitado a recursos a través de HTTP. OAuth2 se basa en la utilización de un token de acceso que el servidor envía al cliente HTTP y con el cual el cliente dispondrá de acceso a los recursos hasta que el token expire después de un tiempo determinado. En el caso de Allrecipes, se ha observado que este token lo proporciona el servidor en las cookies contenidas en la respuesta al solicitar la página principal, tal y como muestra la Ilustración 12:

[21] https://oauth.net/2/

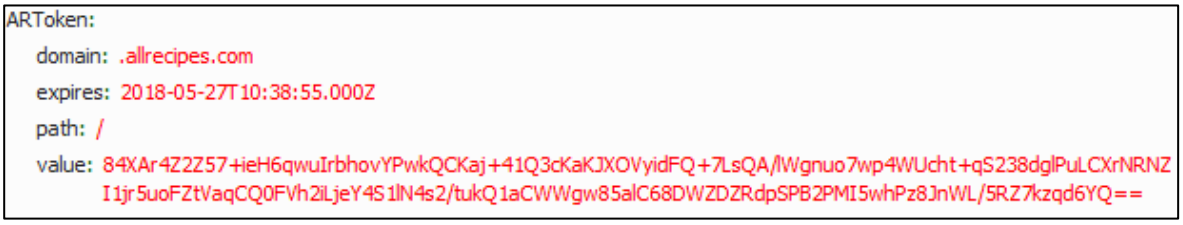

Ilustración 12: Ejemplo de token devuelto por el servidor de Allrecipes

En el ejemplo de un navegador web, este token se utiliza en las cookies de las subsiguientes peticiones, realizadas tanto al servidor web de Allrecipes como a la API, que deberá ser renovado una vez expire. Con ello ya se ha obtenido una manera de disponer de acceso a la API de Allrecipes, tal y como lo realiza un navegador web, y por tanto la posibilidad de solicitar datos a la API.

1.3 Requisitos del software

Una vez se ha definido la disponibilidad y la estructura de la información en la fuente de datos, se han determinado una serie de requisitos que deberá cumplir el conjunto de datos que se desea obtener, atendiendo al aspecto de completitud de los datos que se buscaba en la revisión del estado del arte, así como a las necesidades de posteriores procesos que experimenten con los datos, como por ejemplo la implementación de sistemas de recomendación de cualquier tipo o procesos analíticos.

En este apartado también se detallan los requerimientos asociados al alcance de la ejecución y rendimiento del programa extractor de datos, algunos de los cuales requieren implementar un control riguroso de los tiempos de ejecución y evitar todas aquellas vías a través de las cuales el programa pueda ser restringido por el proveedor de datos.

Requisitos funcionales

RF 1: Historial de usuario

Para cada usuario se debe disponer de un conjunto de 100 recetas que, de manera implícita y sin necesidad de haber calificado numéricamente, reflejen su perfil como comensal o cocinero, y serán extraídas preferentemente de la lista de recetas publicadas, y si no se alcanzaran 100 recetas, se extraerán las restantes de las listas de recetas favoritas y recetas que declaró hacer. Disponer del histórico de cada usuario permitirá incorporar a los procesos analíticos una lista de ítems que los usuarios ya han consumido.

RF 2: Revisiones y valoraciones de usuario

El dataset deberá contener para cada usuario un conjunto de 100 revisiones con sus respectivas valoraciones/calificaciones numéricas asignadas a las recetas. A su vez, se deberá disponer de la información completa de dichas recetas. Las valoraciones numéricas permitirán elaborar sistemas de recomendación basados en ratings, como por ejemplo, Filtrado Colaborativo. Las revisiones, por otro lado, abrirán las puertas al procesamiento de lenguaje natural, permitiendo por ejemplo la inferencia de una valoración numérica o asignación de polaridad a las revisiones.

RF 3: Comunidad de usuario

El proceso de extracción deberá capturar la estructura de red social inherente en la fuente de datos, de tal manera que, para cada usuario, se deberá disponer del perfil, historial y revisiones con valoración de un subconjunto de sus seguidores y amigos (al menos 20 seguidores y 20 amigos), con el fin de poder experimentar con análisis y sistemas de recomendación basados en redes sociales.

RF 4: Revisiones de recetas del usuario

Para cada una de las recetas del historial del usuario se deberá disponer de un conjunto de 100 de revisiones, junto con la valoración numérica asociada, cuyo texto asociado deberá tener una longitud mayor o igual a 50 caracteres. Esto posibilitará disponer de un conjunto de ratings para cada receta, lo cual permitirá experimentar con sistemas de recomendación que utilice los ratings de los diferentes ítems. Además, disponer de un conjunto de revisiones en texto para cada receta ofrecerá la posibilidad de experimentar en otros estudios, como por ejemplo, la sustitución de ingredientes.

RF 5: Información nutricional de receta

La información nutricional de cada receta deberá permanecer en una entidad distinta a su receta asociada en el modelo de datos, con el fin de ofrecer la posibilidad de realizar análisis estadístico a los datos nutricionales sin necesidad de disponer de la información de la receta.

RF 6: Filtrado de usuarios

Se considerará un usuario como válido y apto para incorporarlo al dataset cuando el número de recetas de su historial sea al menos de 30.

Requisitos no funcionales

Escalabilidad:

RNF 1: Paralelismo

El programa extractor de datos debe ser un software ejecutable en distintos entornos o máquinas, con el fin de poder extraer datos en modo paralelo y acelerar así el proceso de obtención de datos.

Operabilidad:

RNF 2: Detección del servidor

El programa debe realizar un control exhaustivo y permitir ajustar los tiempos involucrados en todo el proceso de extracción, intercalando esperas entre los distintos intervalos de ejecución. Esto ayudará a evitar posibles detecciones y denegaciones de acceso por parte del servidor.

RNF 3: Spider traps

El programa debe contemplar la existencia de los tipos de *spider traps* más comunes, que constituyen trampas a nivel de software impuestas por el servidor para evitar programas extractores de datos, y que pueden provocar una finalización inesperada del software, tales como el redireccionamiento HTTP infinito o tamaño excesivo de los documentos HTML.

RNF 4: Formato de los datos

El formato de salida de los datos extraídos deberá ser CSV, delimitado por el carácter barra vertical '|', y un volcado de base de datos relacional SQL.

Integridad:

RNF 5: Integridad de los datos

Los datos extraídos deben resultar completamente íntegros dentro de todo el conjunto de datos, de tal manera que no deberán existir recetas sin un usuario asignado, revisiones sin un autor o receta asignados, o información nutricional e ingredientes sin receta asociada.

Recuperabilidad:

RNF 6: Recuperación del programa

El programa extractor de datos debe contemplar posibles fallos inesperados, tanto por parte del servidor como por parte del entorno donde se ejecuta, o por finalización del programa por parte del usuario. Todos los recursos que esté utilizando, así como el estado de ejecución, deben ser finalizados y salvados correctamente, con el fin de poder recuperar el contexto en ejecuciones posteriores.

Entorno:

RNF 7: Plataforma

El software debe ser ejecutable en un entorno con Máquina Virtual de Java (JVM), para así garantizar que puede ser utilizado prácticamente en cualquier entorno.

2. Diseño

En la fase de diseño se han tomado las decisiones que permiten abordar la extracción de datos desde un aspecto técnico. Estas decisiones incluyen el diseño del modelo de datos que albergará la información extraída y la manera en que los datos van a ser extraídos de la web.

2.1 Modo de extracción

El modo en que se van a extraer los datos ha sido la primera decisión de diseño que se ha tomado. En este proceso se han sopesado las distintas alternativas que existen para extraer los datos, teniendo en cuenta los requisitos funcionales. Dichas alternativas tienen su origen tanto en las diferentes vías de acceso a los datos, como en la propia estructura de la información (deducida en la fase de análisis) de la fuente de datos. La Ilustración 13 muestra un resumen de las fases más relevantes del proceso de extracción.

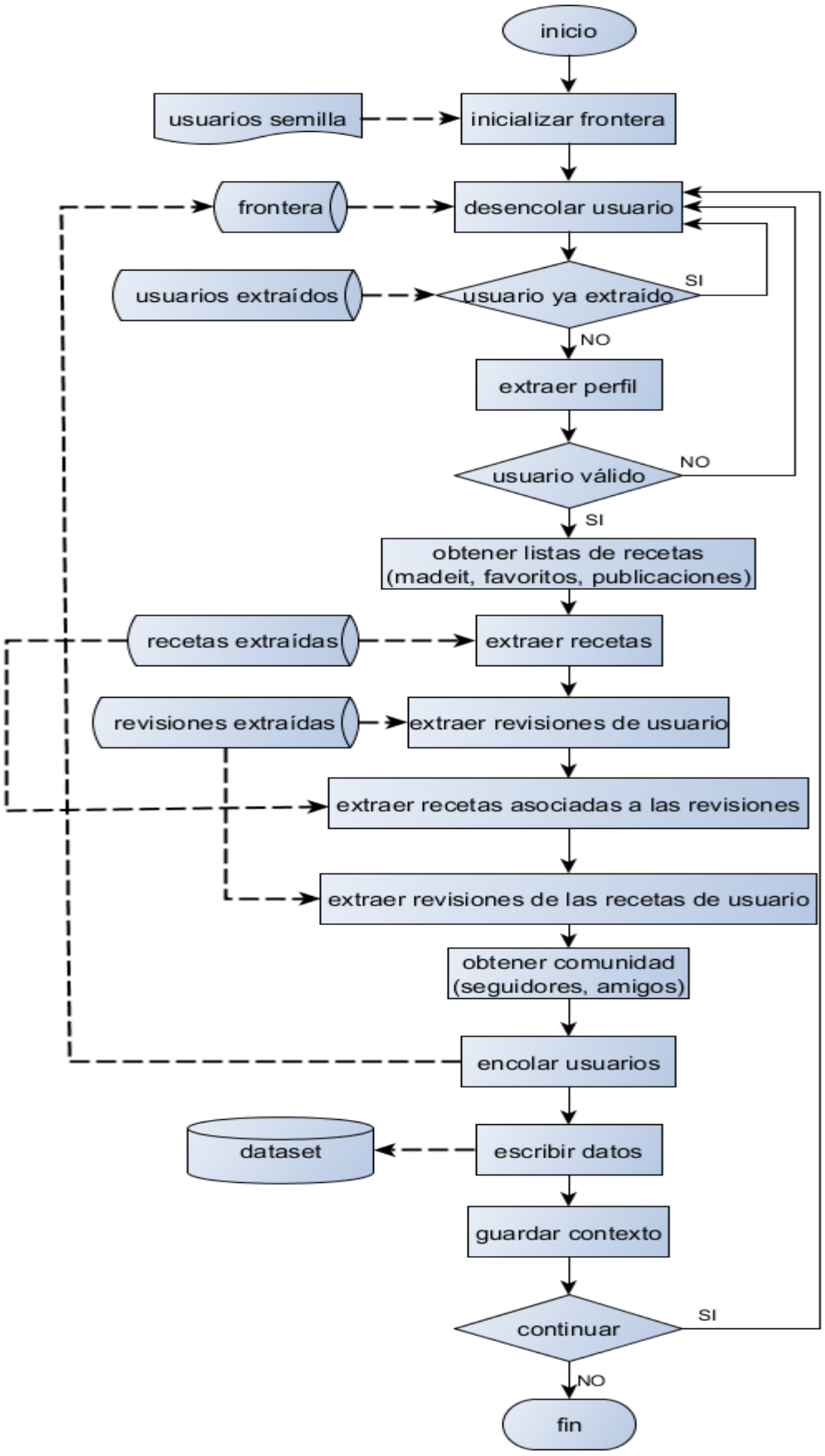

Ilustración 13: Diagrama de flujo del programa extractor de datos

Como se puede observar el proceso guarda cierta semejanza a un proceso de *web crawling*, donde existe una cola de entidades (frontera), de la cual se irá extrayendo el siguiente objeto a procesar. Del mismo modo que en *web crawling*, se deberá mantener un almacén de datos temporal de acceso rápido que albergue las entidades que ya han sido extraídas anteriormente, con el fin de no extraer dos veces el mismo usuario. En este caso también se necesitarán almacenes de datos similares que contengan las recetas y revisiones ya extraídas.

En el procesamiento de cada usuario será extraído en primer lugar su perfil completo, su historial de recetas (recetas publicadas, recetas que declaró hacer y recetas favoritas) y una lista de las revisiones que ha realizado. A su vez, para cada una de estas revisiones se obtendrá la receta asociada y para cada receta extraída del historial del usuario se obtendrá un conjunto de revisiones. Posteriormente se extraerá un subconjunto de sus seguidores y amigos, que serán insertados en la frontera para ser procesados en momentos posteriores. En la fase final el proceso añadirá los datos al modelo de datos que albergue el dataset, guardará el estado de ejecución y decidirá si debe continuar o no con el procesamiento del siguiente usuario. Esta decisión estará basada esencialmente en la cantidad de usuarios, recetas y revisiones que ya han sido extraídos.

El requisito funcional RF3 establece la obligación de capturar los datos (recetas, revisiones y amistad) de la comunidad (amigos y seguidores) de cada usuario, lo cual implica que en la frontera se debe priorizar la extracción de la comunidad de los usuarios. Esto requiere realizar una búsqueda en anchura en el grafo (red social) de los usuarios, y otorgar mayor preferencia a usuarios cuyo usuario objetivo (amigo o seguidor) ha sido extraído en un momento anterior en el tiempo.

El modelo de extracción propuesto cumple con todos los requisitos funcionales expuestos en la fase de análisis, y propone una extracción de datos centrada en los usuarios, obteniendo para cada uno de ellos su historial de recetas, sus revisiones y su red social.

2.2 Modelo de datos

En el diseño del modelo de datos se ha tomado como referencia el diagrama Entidad-Relación y las tablas 4 y 5 mostradas en la fase de análisis. El modelo de datos que se ha elegido es un modelo relacional, que constituye un estándar en bases de datos. Para construir dicho modelo se ha aplicado el algoritmo de transformación de esquema Entidad-Relación a modelo relacional, que junto con el requisito RF6, permiten obtener el modelo de datos mostrado en la Ilustración 14.

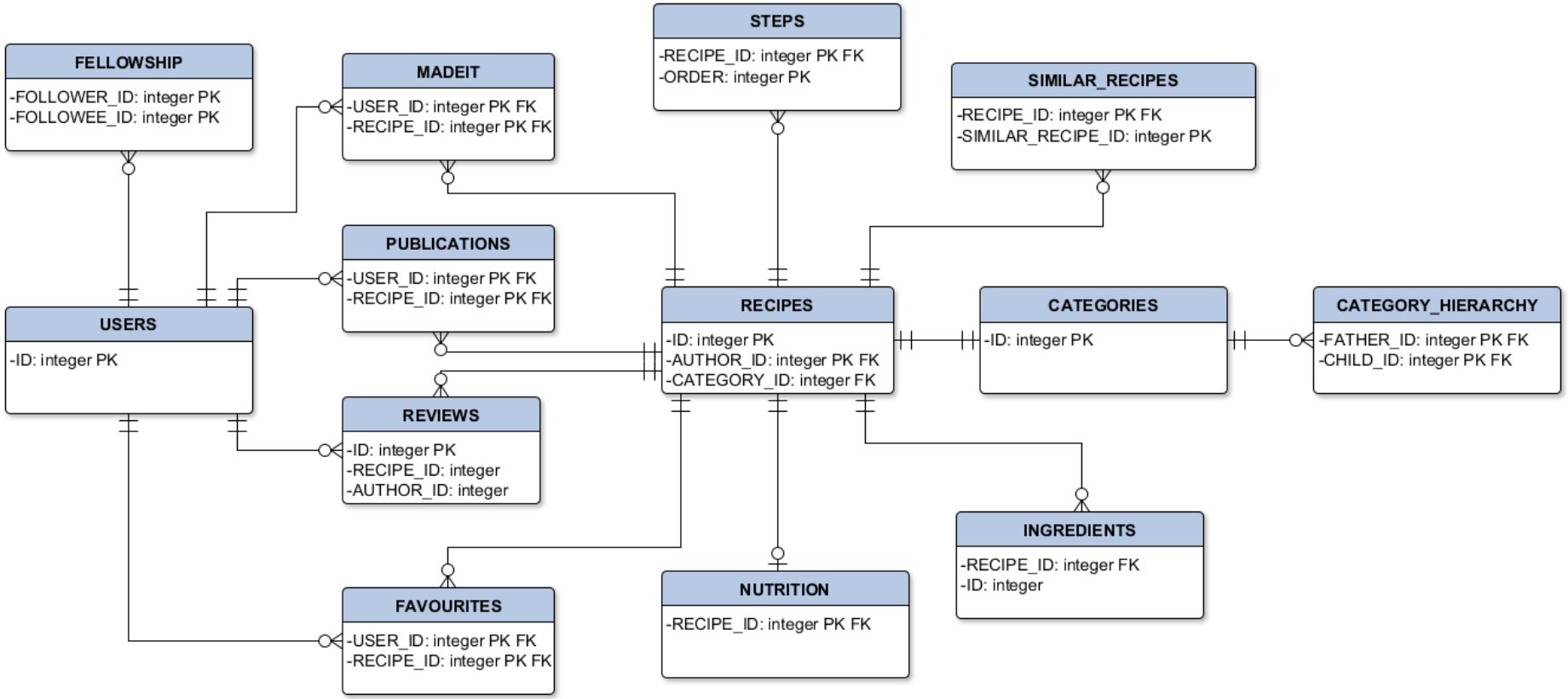

Ilustración 14: Modelo de datos del dataset

Tal y como se puede comprobar, se han incluido únicamente las columnas más relevantes (en el anexo B se puede encontrar la información detallada para cada tabla). Las características más relevantes por destacar están en las siguientes tablas:

- FELLOWSHIP: en otro escenario, ambas columnas (FOLLOWER_ID y FOLLOWEE_ID) serían claves ajenas apuntando a la tabla USERS; sin embargo y según el modelo de extracción planteado, los amigos/seguidores serán extraídos (y por tanto añadidos al catálogo de usuarios de la tabla USERS) en momentos de ejecución posteriores al momento de obtención del usuario objetivo (contenido en la tabla USERS), por lo que no se puede garantizar que los usuarios seguidores/amigos existan en la tabla USERS.
- REVIEWS: en primer lugar tenemos que, al extraer las revisiones de usuario, la receta asociada a dichas revisiones puede no haber sido extraída hasta el momento, y por lo tanto no existir necesariamente en la tabla RECIPES, donde se almacenarán las recetas obtenidas durante el proceso. Además, al extraer las revisiones de receta, tenemos de nuevo que los usuarios autores de dichas revisiones pueden no haber sido extraídos con anterioridad, y por lo tanto no existirán necesariamente en la tabla USERS.
- INGREDIENTS: por último podemos ver que en esta tabla no existe clave primaria, y esto es debido a que un mismo ingrediente puede aparecer más de una vez en una misma receta (utilizado, por ejemplo, en momentos distintos durante la ejecución de la receta).

El modelo de datos propuesto cumple con los requisitos establecidos, y permite albergar el dataset a extraer tanto en ficheros CSV como en una base de datos relacional.

3. Implementación

En este apartado se describirán los detalles de implementación más relevantes del software extractor de datos, tales como los lenguajes de programación y las tecnologías de bases de datos utilizadas.

Frontera

En la fase de diseño se ha establecido realizar una exploración de la red de usuarios basada en búsqueda en anchura. Para satisfacer este requerimiento, la frontera implementa una cola de prioridad, donde la prioridad constituye un valor numérico que indica el momento del tiempo (iteración) en el que el ítem (usuario) ha sido insertado en la cola, tal y como muestran las ilustraciones siguientes:

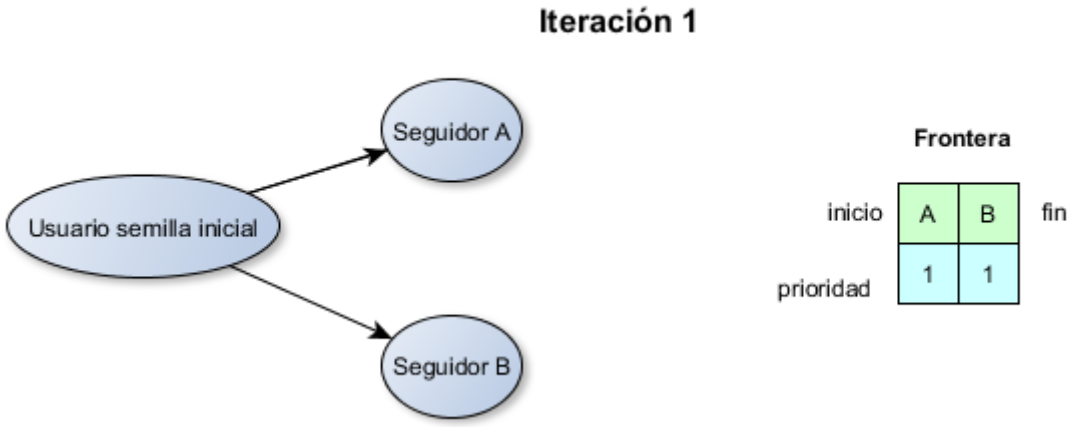

Ilustración 15: Ejemplo de una primera iteración de la cola de prioridad (frontera)

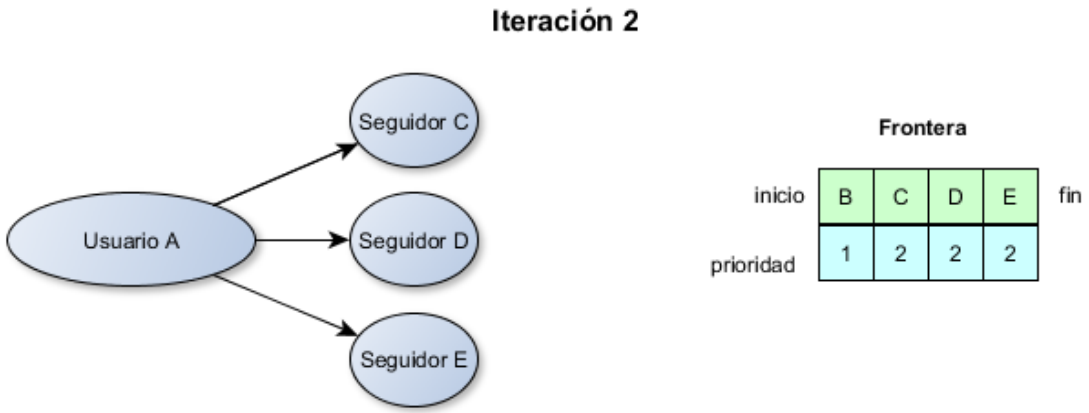

Ilustración 16: Ejemplo de una segunda iteración de la cola de prioridad (frontera)

A partir del funcionamiento de la frontera que muestran las ilustraciones 15 y 16, se puede deducir que no es necesario utilizar una cola de prioridad, ya que sería suficiente con una cola tipo FIFO. Sin embargo, con el fin de poder salvar y recuperar el estado de la cola en cualquier momento de la ejecución, se ha decidido incorporar la prioridad de cada ítem de la cola, así como salvar la última prioridad (iteración) realizada. La frontera a su vez utiliza una tabla de base de datos auxiliar, donde almacena los ítems de todas las iteraciones posteriores a la iteración actual (*i*), dejando así en la propia cola solo los ítems de la iteración *i*, ofreciendo de este modo espacio *ilimitado* a la frontera y evitar desbordar la memoria del programa.

Software de extracción de datos

Lenguajes y librerías

El lenguaje en el que se ha implementado el programa extractor de datos ha sido Scala[22], un lenguaje de programación multi-paradigma ejecutable en JVM, con lo cual se satisface el requisito RNF 7. También se ha utilizado junto con Scala la herramienta SBT[23], que permite gestionar dependencias y realizar tareas de construcción de paquetes Java ejecutables, como por ejemplo los formatos *.jar, .war,* o *.ear*.

En el proceso de extracción se realizan funcionalidades tales como *scraping* y manipulación de objetos JSON, para lo cual se han importado las librerías Jsoup[24] para procesar HTML y *org.json*[25]para JSON. Asimismo también se ha incorporado la librería *scala-csv*[26] para la manipulación de ficheros en formato CSV.

Por otro lado, para la consecución de la versión del dataset generado en formato de volcado de base de datos, se ha incorporado el uso de la tecnología Apache Spark[27], que constituye principalmente un motor de procesamiento en forma de computación distribuida, haciendo uso de la memoria RAM de las máquinas disponibles en un *cluster*. En este contexto se ha utilizado la librería SparkSQL[28], que facilita las tareas de lectura de ficheros en formato columnar (CSV) y su carga en bases de datos SQL.

Bases de datos

El software de extracción de datos hace uso de dos repositorios de datos distintos: por un lado, las tablas auxiliares donde se almacenarán los identificadores unívocos de los datos ya extraídos (recetas, usuarios y revisiones), para lo cual, y teniendo en cuenta el requerimiento RNF 7, se ha elegido Apache Derby[29], que constituye un motor de bases de datos ligeras embebidas en procesos que ejecutan en JVM. Por otro lado, el dataset debe ser generado, además del formato CSV, en una base de datos relacional (acorde al modelo relacional que se ha diseñado). En este caso se ha implementado un proceso independiente que realizará la ingesta desde el formato CSV a una base de datos, en este caso, gestionada a través de PostgreSQL[30].

[22] https://www.scala-lang.org/

[23] https://www.scala-sbt.org/

[24] https://jsoup.org/

[25] https://mvnrepository.com/artifact/org.json/json

[26] https://github.com/tototoshi/scala-csv

[27] https://spark.apache.org/

[28] https://spark.apache.org/sql/

[29] https://db.apache.org/derby/

[30] https://www.postgresql.org/

Gestión de tiempos

Mediante el método de prueba y error se han calculado los tiempos, asociados a cada tarea, que satisfacen el requisito RNF 2. A continuación se adjunta una tabla que resume la cantidad de tiempo de espera tras la finalización de cada tarea relevante del proceso que involucra un contacto con el servidor de la fuente de datos, según el diagrama de flujo del proceso mostrado en la Ilustración 13:

Tarea/Concepto	**Cantidad de tiempo** (segundos)
Extracción de categoría	1
Cada 30 categorías extraídas	30
Extracción del perfil de usuario	0,5
Receta de usuario extraída	3
Cada 50 recetas de usuario extraídas	30
Lista de recetas de usuario extraída	15
Extracción de recetas del usuario	0,5
Extracción de revisiones del usuario	0,5
Receta asociada a revisión de usuario extraída	30
Cada 50 recetas de revisión de usuario extraídas	15
Extracción de recetas de revisiones de usuario	0,5
Revisiones de receta extraída	30
Cada 50 revisiones de receta extraídas	15
Extracción de revisiones de receta de usuario	0,5
Extracción de comunidad del usuario	1
Iteración completa del flujo	8
Cada 20 iteraciones completas	50

Tabla 6: Tiempos de espera utilizados en los programas extractores de datos

Ejecutables

Para cumplir con los requisitos del dataset se han implementado los siguientes programas ejecutables:

Ejecutable 1: *es.eps.uam.tfm.fmendezlopez.AllrrecipesExtractor*

Este programa implementa el proceso principal: la extracción de datos. Debido a la gran cantidad de datos (principalmente usuarios, recetas y revisiones) existentes en Allrecipes, asumimos que el tiempo de ejecución de este programa es ilimitado.

Ejecutable 2: *es.eps.uam.tfm.fmendezlopez.CategoriesExtractor*

Este programa se encarga de extraer los datos de las categorías de Allrecipes. En este ejecutable podemos asumir que el tiempo de ejecución está acotado, ya que en Allrecipes el conjunto de categorías es mucho más reducido que el de recetas, usuarios y revisiones.

Ejecutable 3: *es.eps.uam.tfm.fmendezlopez.CSVtoSQL*

Este último ejecutable permite la ingesta de datos desde los ficheros generados por los dos ejecutables anteriores hacia una base de datos PostgreSQL.

Todo el software de extracción que se ha implementado está disponible a través de GitHub[31].

4. Extracción

La extracción de los datos se ha realizado mediante la ejecución de los programas ejecutables 1 y 2. El tiempo durante el cual el programa 1 ha estado ejecutando es de 25 horas y 20 minutos, mientras que toda la información de las categorías de Allrecipes ha sido extraída por el ejecutable 2 en 2 horas y 20 minutos.

En la Tabla 7 se puede visualizar el detalle de la cantidad de datos que han sido extraídos:

Entidad	Cantidad
Categorías	2588
Recetas	12151
Revisiones/ratings	495210
Usuarios con perfil completo	362
Usuarios con rating	258391

Tabla 7: Cantidad de datos contenidos en el dataset

La estructura del directorio generado que contiene el dataset se puede ver en la Ilustración 17, que contiene un fichero CSV asociado a cada una de las tablas del modelo de datos.

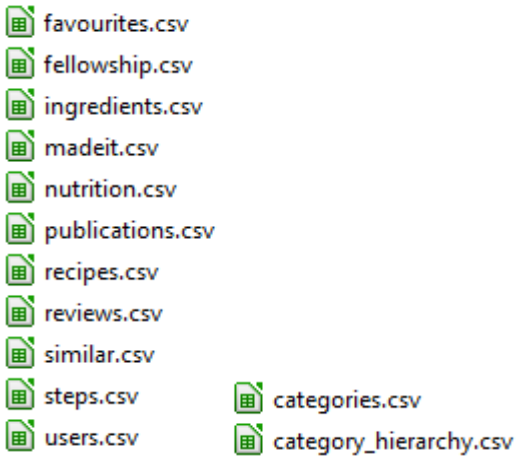

Ilustración 17: Ficheros creados por los programas de extracción de datos

[31] www.github.com/fmendezlopez/TFM

Software de extracción de datos

Capítulo 5

Sistemas de recomendación

La experimentación con el conjunto de datos obtenido mediante el software de extracción de datos con sistemas de recomendación forma parte de los objetivos de este trabajo. A lo largo de este capítulo se proporcionará una visión de los sistemas de recomendación, tanto a nivel general como orientados a la recomendación de recetas de cocina. Asimismo se describirán los dos algoritmos de recomendación que se han utilizado en el dataset obtenido en fases previas, así como los resultados obtenidos. Uno de dichos algoritmos proviene de la revisión de sistemas de recomendación descrita en el capítulo 2, y se ha reproducido para utilizarlo en el dataset generado. El otro de los algoritmos, proviene de la generación de un sistema de recomendación basado en contenido, que utilice principalmente los ingredientes y los datos nutricionales.

1. Motivación

Los sistemas de recomendación constituyen sistemas de filtrado de información cuyo objetivo es intentar predecir la preferencia que un usuario tiene hacia un determinado objeto. Este tipo de sistemas por tanto son aplicables a cualquier ámbito donde exista la necesidad o la posibilidad de sugerir al usuario de cierto software un producto de su interés, teniendo de esta manera sistemas que recomiendan películas, destinos turísticos, productos a la venta en comercio electrónico, o incluso platos de comida. Históricamente el interés por este tipo de sistemas ha incrementado, de la misma manera que evoluciona la era de la digitalización. Asimismo, la competición *The Netflix Prize* [44] ha inspirado a la comunidad a investigar y experimentar en este tipo de sistemas.

En los sistemas de recomendación existen, básicamente, tres enfoques principales, que son los más comunes; el enfoque basado en contenido, que explora las características inherentes a los objetos con los que ha interactuado cada usuario, con el fin realizar sugerencias basadas en el perfil de cada usuario. El método de filtrado colaborativo, por otro lado, explota la similitud entre usuarios, teniendo en cuenta para cada uno sus preferencias sobre los distintos objetos, y asumiendo que usuarios con preferencias semejantes tendrán opiniones similares sobre los diferentes objetos. Por último están los modelos híbridos, que combinan aspectos de los dos métodos anteriores con el fin de mejorar la calidad de este tipo de sistemas.

En el contexto gastronómico se han usado variadas técnicas para recomendar recetas de comida, tal y como se puede observar en la revisión realizada en este tipo de sistemas, detallada en el capítulo 2, tanto basadas en contenido, como filtrado colaborativo, o incluso híbridas. Sin embargo, los más populares han aplicado técnicas basadas en contenido, pudiendo ser porque en este ámbito los objetos a recomendar (recetas) tienen un variado catálogo de características, que son susceptibles de ser explotadas, y que pueden reflejar con cierta solidez las preferencias de usuario.

Siguiendo en esta línea, en este trabajo se han aplicado dos sistemas de recomendación al dataset generado, uno de los cuales ha sido utilizado en uno de los artículos revisados, y el otro es un algoritmo basado en contenido diseñado para el propósito de este trabajo.

2. Diseño de sistemas de recomendación

Tal y como acabamos de mencionar, los sistemas de recomendación empleados en este trabajo son basados en contenido, utilizando como principal característica los ingredientes de las recetas. La arquitectura de software subyacente en ambos algoritmos es la propuesta por Ricci et al. en *Recommender Systems Handbook* [49] para sistemas de recomendación basados en contenido, que consta de los siguientes componentes:

- CONTENT ANALYZER. Este componente se encarga de realizar las tareas adecuadas de preprocesamiento de datos y extracción de características para representar la información en el formato que requieren las técnicas de recomendación a aplicar.
- PROFILE LEARNER. En esta fase se construye el perfil de usuario mediante la combinación de las preferencias de usuario, inducidas a partir de los datos de entrenamiento.
- FILTERING COMPONENT. En este último módulo se aplican métricas de similitud entre el perfil de usuario y cada uno de los elementos del conjunto de datos de test, con el fin de conformar una lista de ítems a recomendar a cada usuario.

El primer sistema de recomendación desarrollado utiliza el algoritmo basado en propagación de ratings, diseñado en el artículo *Recommending Food: Reasoning on Recipes and Ingredients* [18]. Este sistema constituye un algoritmo de aprendizaje supervisado, donde se intenta predecir el rating que un usuario proporcionará a un ítem basándose en los ratings que dicho usuario ha asignado a otros elementos. Para predecir el rating el algoritmo realiza los siguientes procesos:

1. Propagación de rating desde las recetas hacia los ingredientes. En esta primera etapa se calcula, para cada usuario, la valoración promedio asignada a cada uno de los ingredientes que están presentes en aquellas recetas con las que el usuario ha interactuado, obteniendo así un valor que expresa de forma numérica la preferencia de cierto usuario ante cierto ingrediente. Para ello el algoritmo obtiene para cada pareja usuario-ingrediente *(u, i)* una lista de ratings a promediar, que proceden de todos los ratings de aquellas recetas *s* del conjunto R_u, valoradas por *u* y que contienen el ingrediente *i*, tal y como indica la siguiente expresión:

$$\tilde{r}(u,i) = \frac{\sum_{s \in R_u} r(u,s)}{\|R_u\|}$$

   El resultado de este proceso es un perfil de usuario, para cada usuario del conjunto de datos, compuesto por un vector de ratings promediados del usuario para cada ingrediente, con tantas dimensiones como ingredientes distintos estén presentes en las recetas del usuario (R_u).

2. Propagación de ratings desde los ingredientes hacia las recetas. En esta segunda fase, el algoritmo predice los ratings de las recetas que cada usuario aún no ha valorado. Así, para cada pareja usuario-receta *(u, s)*, se promedian los valores de cada ingrediente *i* contenidos en el vector I_u del usuario *u* (calculado en el paso anterior), dividiendo entre el número de ingredientes I_s de la receta objetivo *s*, tal y como muestra la siguiente expresión:

$$\tilde{r}(u, s) = \frac{\sum_{i \in I_u} \tilde{r}(u, i)}{\|I_s\|}$$

Como se puede observar, mediante un método simple que propaga los ratings desde las recetas hacia los ingredientes y a la inversa, se consigue un sistema de recomendación basado en contenido que tiene en cuenta la preferencia del usuario (rating) presente en una información esencial de la receta (ingredientes). Sin embargo y como consecuencia de su simplicidad, este método presenta ciertas carencias como, por ejemplo, la utilización de valores promedios. La media aritmética no captura información referente a la dispersión de los valores (ratings) y de cómo éstos varían. De este modo y en el caso de este algoritmo, un rating promedio de un usuario *U* a un ingrediente *i* que esté, por ejemplo, por debajo de 2 (en una escala de 5), no debería implicar necesariamente que al usuario *U* en general no le gusta el ingrediente *i*. La acción de proporcionar una valoración a una receta tiene en cuenta más propiedades, como pueden ser la presencia o ausencia de ciertos ingredientes, los métodos de cocina utilizados, la gestión de los tiempos de preparación y cocinado, o los valores nutricionales.

Otra característica a destacar es la ausencia de la utilización de la importancia de cada ingrediente, dentro de la colección de ingredientes. Hay ingredientes que por su naturaleza no resultan relevantes y están presentes en muchas recetas, como por ejemplo la sal, el azúcar o el aceite. Según el algoritmo, el valor promedio asignado a este tipo de ingredientes está influyendo con el mismo peso en la predicción del rating de cada receta, para lo cual se debería ponderar de algún modo el peso de cada ingrediente según su rareza en el conjunto de datos, tomando mayor relevancia aquellos ingredientes menos comunes.

Además del algoritmo de propagación de ratings descrito, se ha diseñado otro algoritmo también basado en contenido y, en este caso, no supervisado. Este nuevo algoritmo crea un perfil de usuario compuesto por una serie de vectores que incorporan información nutricional referente al usuario y sus preferencias en ingredientes, a partir de los datos de entrenamiento. Los datos del conjunto de entrenamiento son recetas seleccionadas de las listas de recetas (o historial) del usuario (tablas PUBLICATIONS, FAVOURITES y MADEIT). Los datos del conjunto de test, por otro lado, son recetas a las que el usuario ya ha asignado una valoración o rating, esto es, son recetas pertenecientes a la tabla REVIEWS. De esta manera será posible en pasos posteriores realizar una evaluación de los métodos utilizados, contrastando el rating predicho con la valoración real que el usuario asignó a la receta. Los componentes de este algoritmo se detallan a continuación:

Content Analyzer

Este componente se encarga de realizar las tareas de preprocesamiento y preparación de los datos para poder crear un perfil de usuario (*Profile Learner*) y sugerir contenido (*Filtering Component*). Las distintas tareas que realiza este módulo son las siguientes:

- Filtrado. Las recetas consideradas como válidas son aquellas que tienen información nutricional y lista de ingredientes informadas. Esta tarea implica filtrar todos los ficheros (tablas) del dataset.
- Particionamiento. Tarea que particiona el conjunto de datos filtrado anteriormente en dos subconjuntos: entrenamiento y test. En este caso la estrategia seguida ha sido, tal y como se ha comentado anteriormente, añadir las recetas de usuario de las tablas PUBLICATIONS, FAVOURITES y MADEIT al conjunto de entrenamiento, y las revisiones de usuario al conjunto de test, sin tener en cuenta el cardinal de cada conjunto.
- IDF (*Inverse Document Frequency*). Con el fin de incorporar al sistema de recomendación la penalización de los ingredientes más comunes, se ha añadido al algoritmo el cálculo de IDF para cada ingrediente presente en la colección. Este valor permite expresar la especificidad de un elemento en una colección, y su cálculo ha sido adaptado al caso particular de los ingredientes de la siguiente manera:

$$IDF(i) = \log \frac{|R|}{|R_i|} \quad ,$$

donde R es el número de recetas que existen en el conjunto de datos y R_i es el número de recetas del conjunto R en las que el ingrediente i está presente. De esta manera disponemos de un valor asociado a cada ingrediente que expresa su rareza en la colección total de ingredientes.

Profile Learner

En esta fase el algoritmo construye un perfil para cada usuario existente en los datos de entrenamiento. El perfil de usuario está compuesto por los vectores detallados en la Tabla 8, donde I_u indica el conjunto de ingredientes presentes en las recetas del historial del usuario u, y N hace referencia al conjunto de todos los atributos que describen los datos nutricionales o nutrientes (tabla NUTRITION) que, en este caso, tiene una dimensión de 20 valores:

Vector	Descripción	Tamaño	Valor de cada dimensión
absI	Vector de ingredientes	$\lvert I_u \rvert$	Frecuencia de uso de cada ingrediente
weightI	Vector de ingredientes	$\lvert I_u \rvert$	*absI* * IDF
avgN	Vector de nutrientes	$\lvert N \rvert$	Valor promedio de cada nutriente

Tabla 8: Vectores que componen el perfil de usuario en el algoritmo de recomendación desarrollado

Estos vectores representan cada una de las estrategias utilizadas en el algoritmo de recomendación para poder calcular en la fase *Filtering Component* una medida de similitud entre un perfil de usuario y una receta.

Filtering Component

En esta última etapa se aplican las distintas métricas de similitud a cada revisión (valoración) del conjunto de test. Todas las métricas generadas (Tabla 9) están basadas en similitud coseno, utilizando diferentes combinaciones de los vectores que componen el perfil de usuario, con el fin de realizar comparaciones entre las distintas estrategias utilizadas.

Métrica/estrategia	Utiliza
absI	Vector *absI*
weightI	Vector *weightI*
avgN	Vector *avgN*
absN	Vectores *absI* y *avgN*
weightN	Vectores *weightI* y *avgN*

Tabla 9: Métricas de similitud utilizadas en el algoritmo no supervisado

Cada una de las métricas calculan la similitud coseno entre los vectores que forman el perfil del usuario y la receta objetivo, y en el caso de las métricas *absN* y *weightN* se realiza una suma ponderada de las métricas *absI-avgN* y *weightI-avgN*, respectivamente, otorgando un peso de 0,5 a cada una ellas. Una vez generados los distintos valores de similitud, pertenecientes al intervalo I [0, 1], se ha realizado un proceso de normalización *min-max* para proyectar los valores en el intervalo F [0, 5], utilizado la siguiente expresión:

$$min\_max(x) = \frac{x - min(I)}{max(I) - min(I)} (max(F) - min(F)) + min(F)$$

De esta manera ya tenemos las medidas de similitud expresadas en el mismo intervalo de los ratings, aunque con valor decimal. Para transformar estos valores a números enteros, tal y como está el rating objetivo con el que comparar, se ha aproximado al número entero más próximo. La disponibilidad de esta serie de valores de similitud permitirá en procesos posteriores de evaluación elegir el valor que mejor se adapte a cada tipo de evaluación.

La implementación de ambos sistemas de recomendación se ha orientado de tal modo que puedan ser ejecutados en entornos de procesamiento distribuidos que, aunque para el tamaño del dataset no ha sido necesario, proporciona una escalabilidad adaptable a tamaños mayores del mismo. De esta manera tanto los sistemas de recomendación como los distintos procesos de evaluación han sido implementados utilizando Spark (SparkSQL), que permite el manejo de tablas SQL operando en la memoria del entorno de ejecución.

3. Evaluación

Con el fin de evaluar el rendimiento de los algoritmos implementados se han aplicado las siguientes técnicas de evaluación:

- **Evaluación binaria**. Este primer método de evaluación consiste en deducir si al usuario le ha gustado o no le ha gustado el ítem (receta) en cuestión. Se ha establecido que al usuario le gusta un ítem si el rating asociado es igual o superior a 3 (umbral de decisión), por lo tanto, una predicción constituirá un verdadero positivo si ambos ratings (el real y el predicho) son iguales o superiores a 3.

- **Evaluación por regresión**. Las métricas utilizadas son MAE (*Mean Absolute Error*) y MSE (*Minimum Squared Error*).

- **Evaluación por ranking**. Este método de evaluación es propio de los sistemas de recomendación. Para ello se construyen rankings de diversa longitud, que contienen los ítems con la mejor puntuación predicha (rating) de cada usuario. A su vez, para cada usuario se considera que un ítem es relevante si el rating real que asignó al ítem del ranking es igual o superior a 3. La métrica utilizada en cada ranking ha sido la precisión, que mide la proporción de elementos (recetas) del ranking que son relevantes para el usuario.

En todos y cada uno de los métodos de evaluación se ha utilizado como referencia el rating predicho en el intervalo [0, 5] de números enteros. La diferencia en la estrategia de aprendizaje de los algoritmos que se van a evaluar y comparar (supervisado vs no supervisado) implica que los conjuntos de entrenamiento de ambos deben ser distintos, ya que el algoritmo supervisado utilizará recetas ya valoradas y el no supervisado utilizará recetas del historial del usuario. Consecuentemente y para no dificultar la partición de los datos, el conjunto de test de cada algoritmo también ha sido distinto. Además, y debido a problemas de rendimiento al realizar la ejecución en una computadora convencional, tan solo se ha podido realizar la predicción y evaluación en un conjunto de aproximadamente 1000 predicciones del dataset generado. Los resultados obtenidos para cada método han sido los siguientes:

Evaluación binaria

Strategy	TP	TN	FP	FN	False Positive Rate	Precision	Recall	Specificity
avgN	938	0	9	53	1	0,99	**0,947**	0
absI	354	8	1	637	**0,111**	0,997	0,357	**0,889**
weightI	94	9	0	897	**0**	**1**	0,095	1
absN	830	2	7	161	0,778	**0,992**	0,838	0,222
weightN	801	1	8	190	0,889	0,99	0,808	0,111
propR	924	0	33	3	1	0,966	**0,997**	0

Tabla 10: Resultados de la evaluación binaria en los algoritmos de recomendación desarrollados

En la Tabla 10 se pueden observar las distintas métricas que se han utilizado en las 5 versiones del algoritmo no supervisado (*avgN, absI, weightI, absN, weightN*) y el algoritmo *propR* de propagación de ratings. En primer lugar, vemos que el conjunto de test en general presenta una cantidad baja de valores negativos (TN + FP), lo cual refleja cierta tendencia

por parte de los usuarios a que proporcionen una valoración positiva (mayor o igual que 3). Las estrategias de recomendación *absI* y *weightI* presentan valores altos de falsos negativos (FN), lo cual implica que el *recall* sea muy bajo y por lo tanto, que no resultan eficaces identificando las recetas relevantes para los usuarios. En general todas las estrategias presentan alto valor de *precision*, o lo que es lo mismo, una tasa de aciertos muy alta cuando predicen que al usuario le gustará una receta (predicción positiva). Por otra parte, los algoritmos *absI* y *weightI* han sido capaces de identificar mayor cantidad de casos verdaderos negativos (TN) que el resto, aun teniendo una cantidad de falsos negativos mucho más alta que los demñas los algoritmos. Estos dos algoritmos tienen mayor tendencia a dar una respuesta negativa (es decir, que al usuario no le gusta la receta) y aunque no resulten eficaces identificando los ítems relevantes, sí lo son identificando cuándo un objeto es irrelevante, debido a su alto valor de *specificity*. En el gráfico de la Ilustración 18 se pueden observar la tasa de verdaderos positivos (TPR) de cada algoritmo enfrentada a la tasa de falsos positivos (FPR), utilizando solo el umbral con valor 3 indicado anteriormente, donde los mejores algoritmos se aproximarán en la esquina superior izquierda. Según esta gráfica, el algoritmo de propagación de ratings *propR* es igual de eficaz que un algoritmo de recomendación binario aleatorio, y solo las estrategias
weightN (vector de frecuencia de ingredientes ponderada por IDF combinado con el vector de datos nutricionales) y *avgN* (vector de datos nutricionales) tienen un rendimiento ligeramente superior al aleatorio.

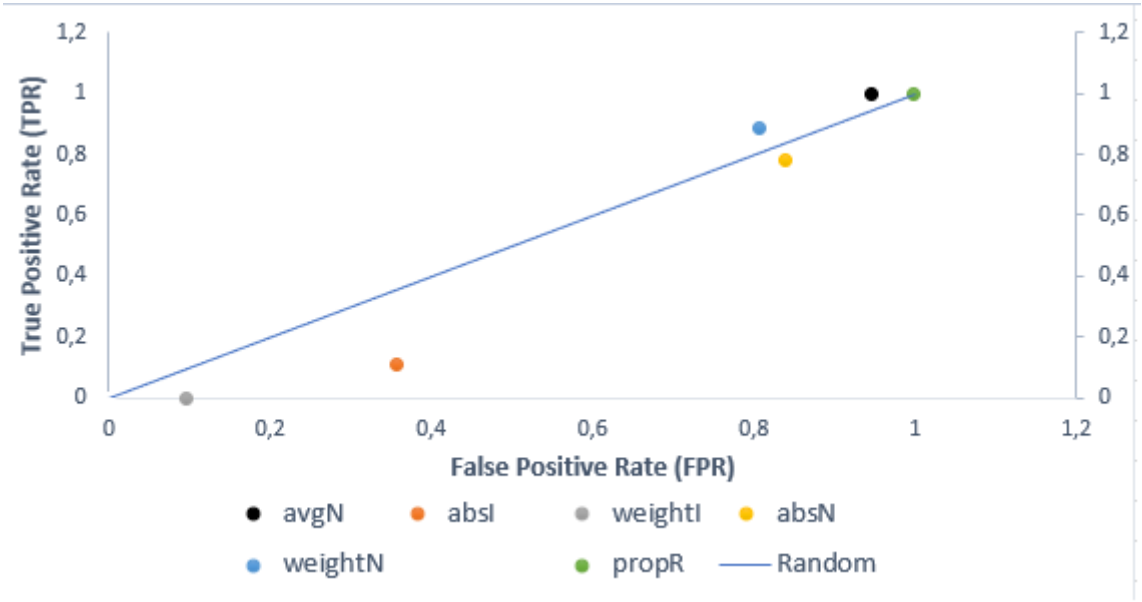

Ilustración 18: Evaluación de TPR vs FPR en los algoritmos desarrollados

Evaluación por regresión

Las estrategias con menor error, entendiendo por error como la diferencia entre el valor del rating predicho y real, han resultado ser *avgN* y *propR*, tal y como se puede ver en la Tabla 11. El algoritmo con menor error promedio es el algoritmo de propagación de ratings (*propR*), y las estrategias *absI* y *weightI*, siendo las más eficaces según la gráfica de la curva ROC (Ilustración 18), son las que mayor error promedio presentan, ya que tenían una alta cantidad de falsos negativos.

Strategy	MSE	MAE
avgN	**1,419**	**0,799**
absI	7,85	2,546
weightI	9,709	2,941
absN	2,887	1,449
weightN	3,354	1,612
propR	**0,908**	**0,54**

Tabla 11: Evaluación de métricas MSE y MAE en los algoritmos desarrollados

Evaluación por ranking

Tal y como se ha comentado anteriormente, la métrica utilizada ha sido P@k (*Precision at k)*, utilizando como longitudes de ranking los valores 5, 10 y 30. En la Tabla 12 podemos ver los resultados de esta evaluación.

Strategy	P@5	P@10	P@30
avgN	**0,988**	**0,994**	0,792
absI	0,965	0,976	0,794
weightI	0,976	0,988	**0,796**
absN	0,976	0,982	0,792
weightN	**0,988**	**0,994**	0,794
propR	0,666	0,42	0,161

Tabla 12: Métrica P@k en los valores 5, 10 y 30 de los algoritmos desarrollados.

Como podemos ver, las estrategias con mayor precisión en las mejores 5 y 10 recomendaciones son *avgN* y *weightN*, alcanzando un valor de 0,988 en P@5. En general, la precisión de ranking en 5 y 10 son superiores en todas las versiones diseñadas para el algoritmo de aprendizaje no supervisado, teniendo valores por encima de 0,96, mientras que el algoritmo de propagación de ratings *propR* tiene una precisión notablemente inferior al resto, decayendo hasta una precisión de 0,161 en P@30. En general, en la métrica *precisión at k* los valores de precisión tienden a decrecer a medida que crece el valor de *k*, lo cual no parece cumplirse con los valores de precisión de 5 y 10 para las estrategias de aprendizaje no supervisado (*avgN*, *absI*, *weightI*, *absN* y *weightN*), donde los valores de P@10 son ligeramente superiores (aunque muy próximos) a los valores de P@5. Sin embargo debemos tener en cuenta que no hay demasiada distancia entre los valores de *k* de estos rankings, al contrario que ocurre como en P@30, y que la evaluación se ha realizado con un conjunto reducido de 1000 predicciones, por lo que es posible que a medida que incremente el número de evaluaciones, el valor de P@10 pase a estar por debajo de P@5.

Capítulo 6

Conclusiones

A lo largo de este trabajo se ha realizado un proceso de ingeniería y ciencia de datos que, motivado por el deseo de poder experimentar con sistemas de recomendación en el contexto gastronómico, ha culminado con el alcance de los objetivos propuestos en este trabajo.

Inicialmente se ha realizado un proceso de recopilación de información en torno a la disponibilidad de conjuntos de datos públicos de gastronomía, encontrando en este proceso la motivación principal de este trabajo: la ausencia de datos públicos de recetas de cocina. Simultáneamente se ha revisado el contexto de los sistemas de recomendación en este ámbito y cuáles han sido las tendencias principales. Además, se han investigado las aportaciones existentes en el contexto de la inferencia de la dificultad de una receta de cocina, evidenciando que una posible solución podría ser la utilización de la entropía de una receta, junto con sus ingredientes y directrices, para deducir su dificultad. En particular uno de los objetivos planteados pretendía enriquecer el dataset generado mediante características adicionales, como por ejemplo, la dificultad de las recetas. Este objetivo no ha sido completado, aunque la investigación realizada y el dataset obtenido han abierto un camino a través del cual poder continuar experimentando. Todo este proceso de revisión ha permitido determinar el resto de los objetivos principales de este trabajo: la generación de un dataset de recetas de cocina y la aplicación de sistemas de recomendación al dataset obtenido.

Para cumplir con los objetivos establecidos, mediante un proceso de ingeniería inversa se ha logrado generar un conjunto de datos, a partir de una red social de Internet que resulta popular en el mundo de la cocina. En el proceso de generación de estos datos se ha capturado la información que se ha considerado más relevante para poder ser utilizada posteriormente en procesos de minería de datos y, en concreto, con sistemas de recomendación. Tal proceso de obtención de los datos se ha orientado a la obtención de perfiles completos de usuarios, comensales o cocineros, que constituyen en definitiva los objetivos principales de los sistemas de recomendación. De esta manera, se ha obtenido, para un conjunto de 362 usuarios una lista de las recetas que, o bien han elaborado, o bien han despertado en ellos cierto grado de preferencia positiva. También se ha conseguido disponer para cada uno de estos usuarios la lista de amigos que tiene en la red social y un conjunto de revisiones, tanto las realizadas por él a otras recetas como las que han realizado otros usuarios a sus recetas. El desarrollo del software encargado de obtener estos datos ha resultado en algunos puntos desafiante, debido en gran medida a que se trata de un software que, entre otras tareas, debe capturar los datos desde una fuente de datos, contemplar los errores que puedan existir en tiempo de ejecución, y guardar el estado de la ejecución en cada momento, todo ello de manera totalmente automática y autónoma. Además, el hecho de readaptar continuamente el software a todas las posibles estructuras y variantes de los ficheros HTML y JSON (retornados por la fuente de datos) ha supuesto una complejidad adicional. A pesar de que el software está diseñado para poder extraer datos por tiempo ilimitado (a menos que cambie la estructura de la fuente de datos), los objetivos de este trabajo no versan sobre la magnitud del dataset, por lo que el subconjunto que se ha obtenido, aunque ha resultado ser reducido

comparado con datasets populares, referentes en la minería de datos como MovieLens[32], ha resultado lo suficientemente extenso como para lograr experimentar con sistemas de recomendación a pequeña escala. Por consiguiente, tanto el dataset generado como el software implementado han abierto las puertas a un mundo de posibilidades en torno a la minería de datos en gastronomía.

Por otro lado, se ha conseguido utilizar este dataset en dos sistemas de recomendación basados en contenido. Uno de ellos ha sido incorporado de la literatura revisada al comienzo de este trabajo, cuya implementación se ha adaptado al contexto de los datos obtenidos. El otro sistema ha sido el resultado de un diseño y elaboración propios de este trabajo, como parte de los objetivos planteados, con distintas versiones de algoritmos que, mediante la combinación de vectores, son capaces de realizar sugerencias de comida basadas en el historial nutricional y uso de ingredientes de cada persona. Aunque ambos sistemas han sido generados utilizado dos tecnologías populares en el ámbito del Big Data (Scala como lenguaje de programación y Spark como entorno de procesamiento distribuido), y aun estando preparados para ser ejecutados en entornos reales de procesamiento a gran escala, los sistemas de recomendación han sido ejecutados de forma local en una computadora personal, lo cual tiene como consecuencia un uso de recursos limitado, pudiendo haber realizado predicciones sobre un conjunto reducido de aproximadamente 1000 instancias. Sin embargo, aun así, el proceso de evaluación ha permitido observar el comportamiento de los sistemas de recomendación desarrollados, determinando que los algoritmos subyacentes creados en este trabajo han alcanzado precisiones por encima del 95% en rankings de las top 5 y 10 sugerencias y que, en general, son eficaces identificando los objetos que *sí* le gustarían al usuario, debido a una baja cantidad de falsos negativos. Además, se ha realizado el diseño para incorporar la rareza de los ingredientes en el momento de realizar las recomendaciones. La incorporación de esta característica ha resultado positiva, ya que hemos visto que la estrategia *weightN*, que combinaba la información nutricional con la frecuencia e IDF de los ingredientes, ha sido una de las estrategias que mejores resultados ha conseguido. Los resultados del análisis realizado han aportado contribuciones a la minería de datos en el contexto gastronómico a la vez que han permitido fijar unos valores *baseline* en el dataset que se ha obtenido en pasos previos.

[32] https://movielens.org/

Bibliografía

[1] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, Recipe recommendation using ingredient networks, *in Proc. 3rd Annu. ACM Web Sci. Conf*. (WebSci'12), Jun. 2012, pp. 298–307.

[2] Ahn, Y., Ahnert, S., Bagrow, J., and Barabasi, A. Flavor network and the principles of food pairing. *Bulletin of the American Physical Society 56* (2011).

[3] Kinouchi, O., Diez-Garcia, R. W., Holanda, A. J., Zambianchi, P. & Roque, A. C. The non-equilibrium nature of culinary evolution. New Journal of Physics 10, 073020 (2008).

[4] Shidochi, Y., Takahashi, T., Ide, I., and Murase, H. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. *In Proc. of the ACM multimedia 2009 workshop on Multimedia for cooking and eating activities*, ACM (2009), 9–14.

[5] A. Hashimoto, N. Mori, T. Funatomi, Y. Yamakata, K. Kakusho, and M. Minoh. Smart kitchen: A user centric cooking support system. *In Proc. 2008 Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 848–854, June 2008.

[6] Jon Malmaud, Earl J. Wagner, Nancy Chang, and Kevin Murphy. 2014. Cooking with semantics. *In Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38.

[7] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, J. Macey, Guide to the carnegie mellon university multimodal activity (cmu-mmac) database, Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, 2009.

[8] Jermsurawong, J. and Habash, N. (2015). Predicting the structure of cooking recipes. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 781–786.

[9] Ueda, M., Takahata, M., and Nakajima, S. User's food preference extraction for personalized cooking recipe recommendation. *Proc. of the Second Workshop on Semantic Personalized Information Management: Retrieval and Recommendatio*n (2011).

[10] I. Weber and P. Achananuparp, Insights from machine-learned diet success prediction, in PSB, 2016.

[11] P. Achananuparp and I. Weber. Extracting food substitutes from food diary via distributional similarity. *arXiv preprint arXiv:1607.08807*, 2016

[12] Wagner, C., Singer, P., and Strohmaier, M. Spatial and Temporal Patterns of Online Food Preferences. In *WWW* (Republic and Canton of Geneva, Switzerland, 2014).

[13] West, R., White, R. W., and Horvitz, E. From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In *WWW* (2013).

Bibliografía

[14] A. Said and A. Bellogín. You are what you eat! tracking health through recipe interactions. *In Proc. of RSWeb'14*, 2014.

[15] M. Harvey, B. Ludwig, and D. Elsweiler. Learning user tastes: a first step to generating healthy meal plans? In *ACM RecSys 2012 LifeStyle Workshop*, 2012.

[16] M. Harvey, B. Ludwig, and D. Elsweiler. You are what you eat: Learning user tastes for rating prediction. In *SPIRE*, pages 153–164, 2013.

[17] T. Kusmierczyk, C. Trattner, and K. Nørvâg. Temporality in online food recipe consumption and production. In *Proc. of WWW'15*, 2015.

[18] J. Freyne and S. Berkovsky. Recommending Food: Reasoning on Recipes and Ingredients. *User Modeling, Adaptation, and Personalization*, pages 381–386, 2010.

[19] Chung, Y. Finding food entity relationships using user-generated data in recipe service. *Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM2012)* (2012) 2611- 2614.

[20] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso. Recipe recognition with large multimodal food dataset. In *ICME Workshops*, pages 1–6, 2015.

[21] Freyne, J., Berkovsky, S., Smith, G.: Recipe Recommendation: Accuracy and Reasoning. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 99–110. Springer, Heidelberg (2011).

[22] Forbes, P., and Zhu, M. Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation. *Proceedings of Recommender Systems* (2011).

[23] C. Wagner, P. Singer, and M. Strohmaier. The nature and evolution of online food preferences. *EPJ Data Science*, 3(1):1–22, 2014.

[24] M. A. El-Dasuky, M. Z. Zashad, T. T. Hamza, A. H. El-Bassiouny, Food Recommendation using ontology and heuristics, Advances Machine Learning Technologies and Applications in Computer and Information Science, 322, 423-429, 2012.

[25] S. Berkovsky, J. Freyne, Group-based recipe recommendations: analysis of data aggregation strategies, in *Proceedings of the 2010 ACM Conference on Recommender Systems*, 2010, pp. 111–118.

[26] Safreno, Doug, Deng, Yongxing. The Recipe Learner. Tech. N.p., 2013.

[27] F.-f. Kuo, C.-T. Li, M.-K. Shan, and S.-y. Lee. Intelligent menu planning: Recommending Set of Recipes by Ingredients. In *Proceedings of the ACM multimedia 2012 workshop on Multimedia for cooking and eating activities - CEA '12*, page 1, New York, New York, USA, 2012. ACM Press.

[28] Aberg, J.: Dealing with malnutrition: A meal planning system for elderly. AAAI, Spring Symposium on Argumentation for Consumers of Health Care, 2006 (2006).

Bibliografía

[29] Erica Greene. 2015. Extracting structured data from recipes using conditional random fields. The New York Times Open Blog.

[30] V. Nedovic, Learning recipe ingredient space using generative probabilistic models, in *Proc. Int. Joint Conf. Artif. Intell. Workshops*, Aug. 2013, pp. 13–18.

[31] Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. 2012. A machine learning approach to recipe text processing. In *Proceedings of Cooking with Computer workshop*.

[32] M. Trevisiol, L. Chiarandini, and R. Baeza-Yates. Buon appetito-recommending personalized menus. In *Proc. of HT'14*, 2014.

[33] Yoko Yamakata, Shinji Imahori, Yuichi Sugiyama, Shinsuke Mori, and Katsumi Tanaka. 2013. Feature extraction and summarization of recipes using flow graph. In *Proceedings of the 5th International Conference on Social Informatics*, LNCS 8238, pages 241–254.

[34] Wesley Tansey, Edward W. Lowe, Jr. y James G. Scott, Diet2Vec: Multi-scale analysis of massive dietary data, In *WWW*.

[35] D. Tasse and N. A. Smith, "SOUR CREAM: Toward semantic processing of recipes," Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-LTI-08-005, May 2008.

[36] Zhang, Q., Hu, R., Mac Namee, B., and Delany, S. Back to the future: Knowledge light case base cookery. In *Proc. of The 9th European Conference on Case-Based Reasoning Workshop* (2008), 15.

[37] Wang, L., Li, Q., Li, N., Dong, G., and Yang, Y. Substructure similarity measurement in chinese recipes. In *WWW*, ACM (2008), 979–988.

[38] Mayumi Ueda, Syungo Asanuma, Yusuke Miyawaki, and Shinsuke Nakajima. 2014. Recipe recommendation method by considering the users preference and ingredient quantity of target recipe. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1.

[39] M. Muller, M. Harvey, D. Elsweiler, and S. Mika. Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *Pervasive Health* 2012, 2012.

[40] Shunsuke Hanai, Hidetsugu Nanba y Akiyo Nadamoto, Clustering for Closely Similar Recipes to Extract Spam Recipes in User-generated Recipe Sites, In *WWW*.

[43] Christoph Trattner y David Elsweiler. Food Recommender Systems, In *WWW* (2017).

[44] James Bennet y Stan Lanning. The Netflix Prize, In *WWW*.

[45] Yehuda Koren. The BellKor Solution to the Netflix Grand Prize, In *WWW* (2009).

[46] Ge, M., Elahi, M., Fernaández-Tobías, I., Ricci, F. & Massimo, D. Using tags and latent factors in a food recommender system. In *WWW* (2015).

Bibliografía

[47] I. Fernández-Tobías and I. Cantador. Exploiting social tags in matrix factorization models for cross-domain collaborative filtering. In *WWW*.

[48] Su-Do Kim1, Yun-Jung Lee1 , Hwan-Gue Cho2 and Seong-Min Yoon. Complexity and Similarity of Recipes based on Entropy Measurement. In *Indian Journal of Science and Technology*, 2016.

[49] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., Recommender Systems Handbook. Springer, 2010.

Anexos

Anexo A. Características de las fuentes de datos exploradas

En este anexo se muestra el estudio realizado para observar en qué medida las distintas fuentes de datos cumplen con los requisitos establecidos.

1. Características de las recetas

A continuación se muestra la presencia de las características de las recetas en las fuentes de datos.

Sitio	Tipo	Valoración media	Valoración escala 5	Like/Me gusta	Dificultad
allrecipes.com	red social	✓	✓	✓	✗
food.com	red social	✓	✓	✗	✗
foodnetwork.com	foro/blog	✓	✓	✗	✓
kraftrecipes.com	foro/blog	✓	✓	✗	✗
bettycrocker.com	foro-red social	✓	✓	✗	✗
yummly.com	red social	✓	✓	✓	✗
myrecipes.com	foro/blog	✓	✓	✗	✗
eatingwell.com	foro/blog	✓	✓	✓	✗
cookpad.com	web	✗	✗	✓	✗
simplyrecipes.com	foro/blog	✗	✗	✗	✗
cooks.com	foro/blog	✓	✓	✗	✗
Puntuaciones		3	3	2	2

Tabla 13: Características de las recetas en las fuentes de datos exploradas (1)

Características de las fuentes de datos exploradas

Sitio	Salubridad	Categorías	#personas que lo hicieron	#revisiones	#valoraciones
allrecipes.com	✓	✓	✓	✓	✓
food.com	✗	✗	✗	✓	✓
foodnetwork.com	✓	✓	✗	✓	✓
kraftrecipes.com	✓	✓	✗	✓	✓
bettycrocker.com	✓	✓	✗	✓	✓
yummly.com	✓	✗	✗	✓	✓
myrecipes.com	✓	✓	✗	✓	✓
eatingwell.com	✓	✓	✗	✓	✓
cookpad.com	✗	✗	✗	✓	✗
simplyrecipes.com	✗	✓	✗	✓	✗
cooks.com	✓	✓	✗	✓	✗
Puntuaciones	2	2	2	2	2

Tabla 14: Características de las recetas en las fuentes de datos exploradas (2)

Sitio	#valoraciones por valor	Descripción receta	Notas de receta	Ingredientes	Ingredientes principales
allrecipes.com	✓	✓	✓	✓	✗
food.com	✗	✓	✗	✓	✗
foodnetwork.com	✗	✗	✗	✓	✓
kraftrecipes.com	✓	✓	✓	✓	✗
bettycrocker.com	✗	✓	✓	✓	✗
yummly.com	✗	✗	✗	✓	✗
myrecipes.com	✓	✓	✗	✓	✗
eatingwell.com	✗	✓	✓	✓	✗
cookpad.com	✗	✓	✗	✓	✗
simplyrecipes.com	✗	✓	✗	✓	✗
cooks.com	✗	✗	✓	✓	✗
Puntuaciones	2	1	1	3	1

Tabla 15: Características de las recetas en las fuentes de datos exploradas (3)

Características de las fuentes de datos exploradas

Sitio	Tipo de plato	Ocasión	Etiquetas	Tiempo total	Tiempo preparación
allrecipes.com	✗	✗	✗	✓	✓
food.com	✗	✗	✗	✓	✓
foodnetwork.com	✓	✓	✓	✓	✓
kraftrecipes.com	✗	✗	✓	✓	✓
bettycrocker.com	✗	✗	✗	✓	✓
yummly.com	✗	✗	✓	✓	✗
myrecipes.com	✗	✗	✗	✓	✓
eatingwell.com	✗	✗	✓	✓	✗
cookpad.com	✗	✗	✗	✓	✗
simplyrecipes.com	✗	✗	✓	✗	✓
cooks.com	✗	✗	✗	✗	✗
Puntuaciones	2	2	2	2	1

Tabla 16: Características de las recetas en las fuentes de datos exploradas (4)

Sitio	Tiempo cocinar	#platos	#calorías	Información nutricional básica
allrecipes.com	✓	✓	✓	✓
food.com	✓	✓	✓	✓
foodnetwork.com	✓	✓	✗	✗
kraftrecipes.com	✗	✓	✓	✓
bettycrocker.com	✗	✓	✓	✓
yummly.com	✗	✓	✓	✓
myrecipes.com	✗	✓	✓	✓
eatingwell.com	✗	✓	✓	✓
cookpad.com	✗	✓	✗	✗
simplyrecipes.com	✓	✓	✗	✗
cooks.com	✗	✗	✗	✗
Puntuaciones	1	1	2	2

Tabla 17: Características de las recetas en las fuentes de datos exploradas (5)

Características de las fuentes de datos exploradas

Sitio	Información nutricional extendida	Pasos	Recetas relacionadas
allrecipes.com	✓	✓	✓
food.com	✓	✓	✗
foodnetwork.com	✗	✓	✓
kraftrecipes.com	✓	✓	✓
bettycrocker.com	✓	✓	✓
yummly.com	✓	✓	✓
myrecipes.com	✓	✓	✓
eatingwell.com	✓	✓	✓
cookpad.com	✗	✓	✓
simplyrecipes.com	✗	✓	✓
cooks.com	✗	✓	✓
Puntuaciones	2	3	1

Tabla 18: Características de las recetas en las fuentes de datos exploradas (6)

2. Características de las revisiones

En este apartado se adjuntan las tablas correspondientes a las revisiones.

Sitio	Revisión	Valoración sin revisión	Fecha revisión
allrecipes.com	✓	✓	✓
food.com	✓	✗	✓
foodnetwork.com	✓	✗	✓
kraftrecipes.com	✓	✓	✓
bettycrocker.com	✓	✓	✓
yummly.com	✓	✗	✓
myrecipes.com	✓	✓	✓
eatingwell.com	✓	✗	✓
cookpad.com	✓	✗	✓
simplyrecipes.com	✓	✗	✓
cooks.com	✓	✗	✓
Puntuaciones	3	2	1

Tabla 19: Características de las revisiones en las fuentes de datos exploradas (1)

Sitio	#likes revisión	Categoría de revisión	Comentario a revisión
allrecipes.com	✓	✓	✗
food.com	✓	✓	✓
foodnetwork.com	✓	✗	✓
kraftrecipes.com	✓	✗	✓
bettycrocker.com	✓	✗	✓
yummly.com	✗	✗	✓
myrecipes.com	✓	✗	✓
eatingwell.com	✗	✗	✗
cookpad.com	✗	✗	✗
simplyrecipes.com	✓	✗	✗
cooks.com	✗	✗	✗
Puntuaciones	2	1	2

Tabla 20: Características de las revisiones en las fuentes de datos exploradas (2)

3. Características de los usuarios

Las siguientes tablas reflejan las puntuaciones y presencia de las características de los usuarios en las fuentes de datos.

Sitio	Recetas favoritas	Colecciones favoritas	Recetas que declaró hacer	Revisiones
allrecipes.com	✓	✓	✓	✓
food.com	✓	✗	✗	✓
foodnetwork.com	✗	✗	✗	✗
kraftrecipes.com	✗	✗	✗	✗
bettycrocker.com	✗	✗	✗	✗
yummly.com	✗	✓	✗	✗
myrecipes.com	✗	✗	✗	✗
eatingwell.com	✗	✗	✗	✗
cookpad.com	✗	✗	✗	✗
simplyrecipes.com	✗	✗	✗	✗
cooks.com	✗	✗	✗	✗
Puntuaciones	3	1	2	3

Tabla 21: Características de los usuarios en las fuentes de datos exploradas (1)

Sitio	Recetas personales	#seguidores	#siguiendo	#favoritos	#recetas	Perfil público
allrecipes.com	✓	✓	✗	✓	✓	✓
food.com	✓	✓	✓	✓	✓	✓
foodnetwork.com	✓	✗	✗	✗	✓	✓
kraftrecipes.com	✗	✗	✗	✗	✗	✗
bettycrocker.com	✗	✗	✗	✗	✗	✓
yummly.com	✓	✗	✗	✗	✓	✓
myrecipes.com	✗	✗	✗	✗	✗	✗
eatingwell.com	✗	✗	✗	✗	✗	✗
cookpad.com	✓	✓	✓	✗	✓	✓
simplyrecipes.com	✗	✗	✗	✗	✗	✓
cooks.com	✗	✗	✗	✗	✗	✗
Puntuaciones	1	2	2	2	2	3

Tabla 22: Características de los usuarios en las fuentes de datos exploradas (2)

4. Puntuaciones de las fuentes de datos

A continuación, en la Tabla 23, se indica el significado que se le atribuye a cada puntuación.

Valor	Significado
1	Prescindible
2	Deseable
3	Imprescindible

Tabla 23: Concepto de cada una de las puntuaciones de las características de las fuentes de datos

Por último tenemos el ránking con las puntuaciones totales asignadas a cada una de las fuentes de datos.

Sitio	Puntuación
allrecipes.com	68
food.com	55
foodnetwork.com	49
kraftrecipes.com	47
bettycrocker.com	46
yummly.com	45
myrecipes.com	44
eatingwell.com	40
cookpad.com	29
simplyrecipes.com	26
cooks.com	24

Tabla 24: Ranking de las puntuaciones asociadas a cada una de las fuentes de datos

Características de las fuentes de datos exploradas

Anexo B. Detalle de los campos del dataset

En este anexo se adjunta el detalle de los campos de cada una de las tablas/ficheros que componen el conjunto de datos generado en la fase de extracción de datos. Para cada campo se indica, además del tipo de dato en SQL y la descripción, el grado de fiabilidad o calidad del dato (3: dato fiable, 2: fiabilidad indeterminada, 1: no fiable). La fiabilidad de los datos se ha determinado únicamente de manera experimental, contrastando algunos ejemplos donde los datos no han resultado ser consistentes. Aunque tan solo algunos atributos aparecen como no fiables, como son las URL, cantidades numéricas o datos personales de los usuarios, resulta relevante tener en cuenta el grado de fiabilidad a la hora de utilizarlos en procesos analíticos.

Detalle de los campos del dataset

1. Tabla “CATEGORIES”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
ID	integer	Identificador único de la categoría	3
NAME	character varying(100)	Nombre asociado a la categoría	3
URL	character varying(200)	URL de Allrecipes de la página web de la categoría	3
COUNT	integer	Número de recetas asociadas a la categoría	1

Tabla 25: Detalle de los campos de la tabla “CATEGORIES” del dataset generado

2. Tabla “CATEGORY_HIERARCHY”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
FATHER_ID	integer	Identificador de la categoría padre	3
CHILD_ID	integer	Identificador de la categoría hijo	3

Tabla 26: Detalle de los campos de la tabla “CATEGORY_HIERARCHY” del dataset generado

3. Tabla “FAVOURITES”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta asignada como favorita	3
USER_ID	integer	Identificador del usuario que ha asignado la receta	3

Tabla 27: Detalle de los campos de la tabla “FAVOURITES” del dataset generado

4. Tabla “FELLOWSHIP”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
FOLLOWER_ID	integer	Identificador del usuario que es seguidor de FOLLOWEE_ID	3
FOLLOWEE_ID	integer	Identificador del usuario que es seguido por FOLLOWER_ID	3

Tabla 28: Detalle de los campos de la tabla “FELLOWSHIP” del dataset generado

5. Tabla “INGREDIENTS”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta	3
ID	integer	Identificador del ingrediente	3
TEXT	text	Descripción del ingrediente que proporciona el autor	3
AMOUNT	double precision	Cantidad asociada al ingrediente en miligramos (mg)	3

Tabla 29: Detalle de los campos de la tabla “INGREDIENTS” del dataset generado

6. Tabla “MADEIT”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta declarada como “realizada”	3
USER_ID	integer	Identificador del usuario que ha declarado realizar la receta	3

Tabla 30: Detalle de los campos de la tabla “MADEIT” del dataset generado

7. Tabla “REVIEWS”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
ID	integer	Identificador único de la revisión	3
RECIPE_ID	integer	Identificador de la receta a la que está asociada la revisión	3
AUTHOR_ID	integer	Identificador del autor de la revisión	3
RATING	integer	Valoración numérica proporcionada en la revisión	3
TEXT	text	Texto proporcionado en la revisión	3
DATE	date	Fecha de realización de la revisión en formato YYYY-MM-DD	3
HELPFUL_COUNT	integer	Número de veces que la revisión ha sido considerada como “útil”	3

Tabla 31: Detalle de los campos de la tabla “REVIEWS” del dataset generado

8. Tabla “NUTRITION”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	Integer	Identificador de la receta	3
CALCIUM	double precision	Cantidad asociada a este concepto (mg)	3
CALORIES	double precision	Cantidad asociada a este concepto (kcal)	3
CALORIES_FROM_FAT	double precision	Cantidad asociada a este concepto (kcal)	3
CARBOHYDRATES	double precision	Cantidad asociada a este concepto (g)	3
CHOLESTEROL	double precision	Cantidad asociada a este concepto (mg)	3
FAT	double precision	Cantidad asociada a este concepto (g)	3
FIBER	double precision	Cantidad asociada a este concepto (g)	3
FOLATE	double precision	Cantidad asociada a este concepto (mcg)	3
IRON	double precision	Cantidad asociada a este concepto (mg)	3
MAGNESIUM	double precision	Cantidad asociada a este concepto (mg)	3
NIACIN	double precision	Cantidad asociada a este concepto (mg)	3
POTASSIUM	double precision	Cantidad asociada a este concepto (mg)	3
PROTEIN	double precision	Cantidad asociada a este concepto (g)	3
SATURATED_FAT	double precision	Cantidad asociada a este concepto (g)	3
SODIUM	double precision	Cantidad asociada a este concepto (mg)	3
SUGARS	double precision	Cantidad asociada a este concepto (g)	3
THIAMIN	double precision	Cantidad asociada a este concepto (mg)	3
VITAMIN_A	double precision	Cantidad asociada a este concepto (IU)	3
VITAMIN_B6	double precision	Cantidad asociada a este concepto (mg)	3
VITAMIN_C	double precision	Cantidad asociada a este concepto (mg)	3

Tabla 32: Detalle de los campos de la tabla “NUTRITION” del dataset generado

9. Tabla “PUBLICATIONS”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
ID	integer	Identificador de la receta publicada por USER_ID	3
USER_ID	integer	Identificador del usuario que ha publicado la receta USER_ID	3

Tabla 33: Detalle de los campos de la tabla “PUBLICATIONS” del dataset generado

10. Tabla “RECIPES”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta	3
CATEGORY_ID	integer	Identificador de la categoría asociada a la receta	3
TITLE	text	Título de la receta	3
RATING	double precision	Rating promedio de la receta	3
RATING_COUNT	integer	Número de valoraciones que ha recibido la receta	3
REVIEW_COUNT	integer	Número de revisiones que ha recibido la receta	3
MADEIT_COUNT	integer	Número de usuarios que declararon realizar la receta	3
DESCRIPTION	text	Descripción de la receta	3
SERVING_COUNT	integer	Número de raciones	3
PREP_TIME	integer	Tiempo de preparación	3
COOK_TIME	integer	Tiempo de cocinado	3
TOTAL_TIME	integer	Tiempo total de realización de la receta	3
WEB_URL	character varying(200)	URL del sitio web de la receta en Allrecipes	2
COOK_NOTE	text	Nota del autor de la receta	3
EDITOR_NOTE	text	Nota del autor de la receta	3
TIP	text	Consejo del autor de la receta	3
CHEF_NOTE	text	Nota del autor de la receta	3

Tabla 34: Detalle de los campos de la tabla “RECIPES” del dataset generado

11. Tabla “SIMILAR_RECIPES”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta principal	3
SIMILAR_RECIPE_ID	integer	Identificador de la receta considerada similar a la receta principal	3

Tabla 35: Detalle de los campos de la tabla “SIMILAR_RECIPES” del dataset generado

12. Tabla “STEPS”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
RECIPE_ID	integer	Identificador de la receta	3
ORDER	integer	Número entero que refleja el orden de la instrucción dentro del procedimiento de la receta	3
TEXT	text	Texto asociado a la instrucción	3

Tabla 36: Detalle de los campos de la tabla “STEPS” del dataset generado

Detalle de los campos del dataset

13. Tabla “USERS”

Atributo	Tipo de dato SQL	Descripción	Fiabilidad
ID	integer	Identificador del usuario	3
NAME	character varying(50)	Nombre del usuario	3
CITY	character varying(50)	Ciudad	2
REGION	character varying(50)	Región geográfica	2
COUNTRY	character varying(25)	País	2
HANDLE	character varying(25)	Nombre del usuario	3
URL	character varying(200)	URL de la página web del usuario en Allrecipes	2
FOLLOWER_COUNT	integer	Número total de seguidores del usuario	3
FOLLOWING_COUNT	integer	Número total de amigos del usuario	3
FAV_COUNT	integer	Cantidad total de recetas favoritas del usuario	1
MADEIT_COUNT	integer	Cantidad total de recetas que el usuario ha declarado realizar	1
RATING_COUNT	integer	Cantidad total de valoraciones del usuario	1
RECIPE_COUNT	integer	Cantidad total de publicaciones del usuario	1
REVIEW_COUNT	integer	Cantidad total de revisiones realizadas por el usuario	1

Tabla 37: Detalle de los campos de la tabla “USERS” del dataset generado

Detalle de los campos del dataset